UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

ANÁLISE DO PERFIL DOS COLÉGIOS MILITARES BASEADO EM DADOS DE RENDIMENTOS DE ENSINO

DISSERTAÇÃO DE MESTRADO

Fernando Monteiro Silva

Santa Maria, RS, Brasil

2005

ANÁLISE DO PERFIL DOS COLÉGIOS MILITARES BASEADO EM DADOS DE RENDIMENTOS DE ENSINO

por

Fernando Monteiro Silva

Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Engenharia de Produção, Área de Concentração em Qualidade e Produtividade, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestre em Engenharia de Produção.**

Orientador: Prof. Adriano Mendonça Souza

Santa Maria, RS, Brasil
2005

Endereço: Rua Padre Gabriel Bolzan, n. 1777 – casa 113, Bairro São José, Santa Maria, RS, 97095-500
Fone (55) 3222-9521; Cel (55) 9107-4500; End. Eletr: monteiro@dep.ensino.eb.br

Universidade Federal de Santa Maria
Centro de Tecnologia
Programa de Pós-Graduação em Engenharia de Produção

A Comissão Examinadora, abaixo assinada,
aprova a Dissertação de Mestrado

ANÁLISE DO PERFIL DOS COLÉGIOS MILITARES BASEADO EM DADOS DE RENDIMENTOS DE ENSINO

elaborada por
Fernando Monteiro Silva

como requisito parcial para obtenção do grau de
Mestre em Engenharia de Produção

COMISÃO EXAMINADORA:

Adriano Mendonça Souza, Dr. (UFSM)
(Presidente/Orientador)

Robert Wayne Samohyl, Ph. D. (UFSC)

Luis Felipe Dias Lopes, **Dr.** (UFSM)

Santa Maria, 25 de maio de 2005.

Este trabalho é dedicado aos meus pais.

AGRADECIMENTOS

A Deus, por me proporcionar a oportunidade de desenvolver este trabalho em instituições de tão distinto valor como o Exército Brasileiro e a Universidade Federal de Santa Maria.

Ao professor Adriano Mendonça Souza, um amigo de estimado caráter, educação e conhecimento, pela disposição, paciência e empenho dedicados neste trabalho.

À Universidade Federal de Santa Maria pela oportunidade concedida.

Ao Coronel Guelfi, Coronel Paulo Gil e professor Jorge Costa pelo reconhecimento de meu trabalho junto ao ensino no Departamento de Ensino e Pesquisa do Exército Brasileiro.

Aos Comandantes do Colégio Militar de Santa Maria, Coronel Afonso José Cruz Auler, Coronel Paulo Roberto Santiago Ferreira e Tenente Coronel Herventon Francisco de Assis Maria, por permitirem que eu desempenhe a função de analista e desenvolvedor de sistemas nesta unidade de ensino militar, e pela disposição para realização deste curso.

Ao Tenente Coronel Losada, Major Figueiredo, Capitão Epitácio, Capitão Ribas e Tenente Baldissera, pelo apoio e empenho dedicados no desenvolvimento do Sistema de Gestão Escolar.

Aos colegas do Colégio Militar de Santa Maria e da Equipe de Desenvolvimento do Departamento de Ensino e Pesquisa, por suas colaborações espontâneas que ajudaram muito na minha formação.

A todos os professores do PPGEP, colegas e amigos que contribuíram para a realização deste trabalho.

All Knowledge is, in final analysis, History.
All sciences are, in the abstract, Mathematics.
All judgements are, in their rationale, Statistics.
Radhakrishna Rao

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal de Santa Maria

ANÁLISE DO PERFIL DOS COLÉGIOS MILITARES BASEADO EM DADOS DE RENDIMENTOS DE ENSINO

AUTOR: FERNANDO MONTEIRO SILVA
ORIENTADOR: ADRIANO MENDONÇA SOUZA
Data e Local da Defesa: Santa Maria, 25 de maio de 2005.

Com o objetivo de determinar o perfil dos alunos e dos Colégios Militares, aplicam-se técnicas estatísticas multivariadas em dados dos alunos, disponibilizando assim subsídios para a tomada de decisões da administração. Realiza-se, primeiramente, a análise descritiva dos dados de quatro Colégios Militares, segundo um cruzamento de variáveis de rendimento escolar. Aplica-se análise multivariada em alguns indicadores de ensino, utilizando-se análise de aglomeração, componentes principais e discriminante. Pode-se identificar o número de alunos com rendimento baixo nas classes de origem social e compará-los entre as escolas. Utilizando graus das disciplinas, verificam-se agrupamentos que representam os atributos da área psicomotora/afetiva e ciências/cognitivas. Mostra-se um modelo formado pelas disciplinas mais representativas em um ano letivo comparadas com a necessidade de recuperação, classificando um aluno novo neste sistema. As técnicas aplicadas em indicadores de ensino mostram-se adequadas para a verificação da qualidade, pois, obedecendo a natureza multivariada, pode-se extrair informações relevantes utilizando-se diferentes casos e variáveis. A procura de conhecimento em bancos de dados tem se destacado como uma atividade desejável para descobrir e compreender padrões ocultos nas instituições, os quais nem sempre são visíveis através da simples observação. Desta forma, busca-se aumentar a competência e a criatividade nas instituições públicas, visando a organização e gestão de sistemas de qualidade, através do uso de metodologia para mostrar o desempenho comparativo entre as escolas e entre os próprios alunos.

Palavras-chave: Ensino, militar, estatística multivariada, mineração de dados.

ABSTRACT

Mastership Dissertation
Post-graduation in Engineering Production
Federal University of Santa Maria

ANALYSIS OF THE PROFILES OF MILITAR SCHOOLS BASED IN TEACHING EFFICIENCY DATA

AUTHOR: FERNANDO MONTEIRO SILVA
SUPERVISOR: ADRIANO MENDONÇA SOUZA
Date and Local: Santa Maria, May 25th 2005.

In order to determine the profile of pupils and the Militar Schools, multivariate statistical techniques on the students' data are applied, providing therefore, aids for the administration decision takings. First, a data descriptive analysis of four Militar Schools was carried out according to a cross of school efficiency variants. The multivariate analysis is applied in some teaching indicators by using agglomeration analysis as well as principal components and discriminatory. It is possible to identify the number of students presenting low school grades in their social classes and to compare them among the schools. Once grades in disciplines are used, sets that represent attributes in psychomotor/affective and scientific/cognitive areas are verified. A model formed by the most representative disciplines in the school year, compared to the need of recovery, is shown, classifying a new student in this system. The techniques applied in teaching indicators seem to be suitable to check the quality because when the multivariate nature is obeyed, relevant information can be taken by using different variants and cases. The search for knowledge in data banks has been distinguished as a desirable activity to find out and to understand hidden standards in institutions once they are not always seen through ordinary observation. Thus, it is aimed to enhance the competence and creativity in public institutions looking for organization and quality system management by the usage of methodology in order to show the comparative performance among schools and their own pupils.

Key words: teaching, military, multivariate statistic, data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AA – Análise de Aglomeração

ACP - Análise de Componentes Principais

ADE - Análise Descritiva

ADI - Análise de Discriminante

AM - Análise Multivariada

Bio - Biologia

CMBH - Colégio Militar de Belo Horizonte

CMC - Colégio Militar de Curitiba

CMRJ - Colégio Militar do Rio de Janeiro

CMSM - Colégio Militar de Santa Maria

CP – Componentes Principais

DEP - Departamento de Ensino e Pesquisa

DEPA - Diretoria de Ensino Preparatório e Assistencial

DM - *Data Mining*

EF - Educação Física

Fis - Física

Geo - Geografia

GrauComp - Grau de Comportamento

Hist - História

KDD - *Knowledge Discovey in Databases*

LEM - Língua Estrangeira Moderna

Lit - Literatura

Mat - Metemática

MGS - Média Geral da Série

NPCE - Normas de Planejamento e Conduta do Ensino

PEG - Programa de Excelência Gerencial

Port - Língua Portuguesa

PPerd - Pontos Perdidos

PR - Prova de Recuperação

SCMB - Sistema Colégio Militar do Brasil

SGE - Sistema de Gestão Escolar

TA - Tipo do Amparo

Qui - Química

SUMÁRIO

INTRODUÇÃO

As organizações estão fazendo aquisição de novas tecnologias, implementando processos em rotinas informatizadas, investindo em conectividade e integrando bancos de dados sempre com a preocupação de ter a informação como suporte à decisão. Essas inovações deixam de ser apenas para apoio aos processos produtivos, tornando-se parte integrante deles, muitas vezes redefinindo a maneira de se realizar as atividades. Estamos passando à sociedade do conhecimento, onde a mercadoria mais valiosa é a informação. O fato é a transição de uma era para outra e, junto com ela, as mudanças culturais e adaptações às novas ferramentas tecnológicas.

As mudanças advindas com a sociedade da informação provocaram alterações nos hábitos de uso da informação no dia-a-dia das pessoas, impulsionando as organizações para a busca de uma modernização e agilização dos serviços prestados.

As duas últimas décadas acompanharam um aumento na quantia de informações que são armazenadas em formato eletrônico. Silberschatz (2003) afirma que esta acumulação de dados acontece a uma taxa que dobra a cada vinte meses e o tamanho e número de bancos de dados estão aumentando até mais rapidamente. Tendo concentrado muita atenção no armazenamento de dados, o problema passou a ser o que fazer com estes recursos valiosos. Dados crus têm raramente benefícios diretos. Seu valor verdadeiro é verificado na habilidade de extrair informações confiáveis e úteis.

A procura de conhecimento em bancos de dados tem se destacado como uma atividade desejável para descobrir e compreender os padrões ocultos nas instituições. Esses padrões são usados em modelos que prevêem os comportamentos individuais e dinâmicos com alta precisão.

Um desafio enfrentado hoje pelo ensino é a previsão da trajetória dos alunos. Quais precisarão de assistência adicional para aprovação? Como aumentar a aprovação sem diminuir o conteúdo programático? Quais alunos têm maior probabilidade de ingressar em agremiações e atividades extracurriculares? Enquanto isso, questões tradicionais, tais como o gerenciamento de matrículas, históricos e boletins, continuam a exercer pressão sobre as instituições para que procurem soluções novas e mais rápidas. As escolas podem tratar melhor desses desafios relativos ao ensino através da busca de conhecimento em bases de dados estruturadas.

Melhorar a gestão do ensino significa qualificar o seu produto. É necessário mensurar estatisticamente as múltiplas variáveis que representam os fatores de qualidade de ensino e representá-las numa dimensão compreensível para o administrador.

Segundo Ferraudo (2005), através da tecnologia dos computadores, a quantidade de informação que se pode tratar e armazenar é muito grande, complexa e variada. Na posse de uma enorme quantidade de informações, a questão que surge é naturalmente como interpretá-las e, obedecendo à natureza multivariada, como extrair informação relevante.

A mineração de dados fornece um método automático para se descobrir padrões em dados, sem a limitação de uma análise baseada apenas na intuição. Segundo Berry (1997), as técnicas de mineração mostram-nos como alcançar rápido e facilmente a solução nos negócios, a qual adormece nos nossos sistemas de informação.

O principal motivo que tem levado os administradores a investir nessa tecnologia tem sido a obtenção de uma melhor visão sobre a extensão da base de dados e a revelação de relações implícitas de padrões entre os dados que nem sempre são visíveis através da simples observação. Com sua utilização centrada na busca de relações que permitam identificar novas

informações sobre uma determinada base de conhecimento, a mineração de dados vem se tornando uma ferramenta fundamental para a tomada de decisões.

Segundo Klosgen (2002), *Knowledge Discovey in Databases* (KDD) – Descoberta de Conhecimento em Bases de Dados - oferece métodos informáticos para a obtenção de conhecimento implícito em grandes conjuntos de dados. Data minig (DM) é a etapa de extração no processo de KDD, é a exploração e análise, de maneira automática ou semiautomática, de uma ampla quantidade de dados com o objetivo de descobrir novos conceitos, correlações ou modelos. DM relaciona-se com a análise de dados e o uso de ferramentas computacionais na busca de características, regras e regularidades em um grande conjunto de dados.

As ferramentas de exploração de dados combinam funções de estatística, ciências da computação e recursos de inteligência artificial. A escolha da combinação de técnicas, para serem aplicadas numa particular situação, depende da natureza das tarefas de pesquisa e da natureza dos dados avaliados. Classificação, estimação, predição, agrupamento por afinidade, clusterização e descrição são algumas das tarefas que caracterizam uma exploração de dados.

Segundo Louzada Neto (2000), DM parece não ser novo para muitos estatísticos e econometristas, e tem sido utilizado para descrever o processo de pesquisa de conjunto de dados, na esperança de identificar comportamentos ou características comuns.

A integração entre a mineração e a base de dados pode gerar conhecimento através da procura de padrões de relacionamento e regras associativas de conteúdo. Nesse contexto, um sistema de banco de dados deve permitir que as informações úteis ao desempenho da organização sejam associadas aos acontecimentos de um universo maior, criando vínculos entre os dados internos e as informações externas que se têm sobre a concorrência e os cenários de negócio que se apresentam ao longo do tempo.

1.1 Tema da pesquisa

O tema da pesquisa refere-se a aplicação de técnicas estatísticas multivariadas na exploração do banco de dados dos Colégios Militares, utilizando informações de desempenho e cadastro dos alunos dos Colégios Militares do Brasil, com a intenção de traçar um perfil da escola, e dos alunos, e predizer o rendimento desses na área de ensino.

1.2 Justificativa

A falta de uma ferramenta para demonstração do desempenho comparativo entre diferentes escolas e a necessidade de uma melhor quantificação do evento avaliativo, que normalizam e conferem um caráter objetivo ao fator desempenho escolar para a tomada de decisão dos administradores do ensino, é o que determina a elaboração deste estudo.

Assim, busca-se aumentar a competência e a criatividade nas instituições públicas, visando à organização e gestão de sistemas de qualidade, através do uso de metodologia eficaz para mostrar o desempenho comparativo entre as escolas e entre os próprios alunos.

1.3 Objetivos

Neste item, apresentam-se os objetivos que nortearão este trabalho.

1.3.1 Objetivo geral

Determinar o perfil dos alunos e dos Colégios Militares, utilizando técnicas estatísticas multivariadas aplicadas no rendimento dos alunos, disponibilizando assim subsídios para a tomada de decisões da administração.

1.3.2 Objetivos específicos

- Identificar padrões, classificações e previsões baseadas em técnicas estatísticas, utilizando ferramentas tecnológicas;
- Estudar a relação entre as variáveis em estudo;
- Detalhar as técnicas estatísticas aplicadas na exploração de dados.

1.4 Metodologia

Para que os objetivos desta pesquisa fossem alcançados, uma revisão de literatura foi realizada, com interesse voltado para a descoberta de conhecimento em bases de dados, focando a gestão da qualidade do ensino militar no Brasil.

Utiliza-se primeiramente uma Análise Descritiva (ADE) dos dados de quatro Colégios Militares, fazendo-se um cruzamento de variáveis de rendimento escolar. Depois, parte-se para a Análise Multivariada (AM) de alguns indicadores de ensino.

Para se realizar uma exploração de dados com várias variáveis foi necessário o estudo de algumas técnicas multivariadas como a Análise de Aglomeração (AA), Análise de Componentes Principais (ACP), e Análise Discriminante (ADI). Trata-se de métodos de sumarização de conjuntos de dados, os quais apresentam distintas características.

Também foi necessária a realização de estudos de casos que revelaram o comportamento dos bancos de dados.

1.5 Delimitação do tema

Esta pesquisa constitui-se de um conjunto de técnicas multivariadas aplicadas em dados de quatro Colégios Militares que são: Colégio Militar do Rio de Janeiro (CMRJ),

Colégio Militar de Santa Maria (CMSM), Colégio Militar de Curitiba (CMC) e Colégio Militar de Belo Horizonte (CMBH).

Estes Colégios utilizam o Sistema de Gestão Escolar (SGE), programa de computador desenvolvido pelo Departamento de Ensino e Pesquisa, o qual objetiva atender as necessidades da área de ensino e militar, o qual atende as necessidades de cadastramento de pessoal, controle de matrículas e graus, controle disciplinar, controle de biblioteca e saúde.

As bases de dados (*Oracle* e *PostgreSQL*) possuem a mesma estrutura (esquema), o que facilitou a modelagem dos projetos definidos neste trabalho.

Utilizam-se dados da área de ensino como graus, rendimentos, médias finais e dados de cadastro.

1.6 Importância do trabalho

O trabalho torna-se importante na medida em que há necessidade de se conhecer tanto o perfil da escola quanto o dos alunos que a freqüentam. Pois, desta forma, o comando das instituições de ensino pode tomar decisões em relação ao programa de ensino, práticas pedagógicas e, até mesmo, conhecer a vocação do local onde a escola se encontra.

Este trabalho tem contribuições significativas no âmbito administrativo do ensino, pois, utilizando ferramentas e técnicas estatísticas, disponibilizam-se informações e definem-se padrões, os quais podem também servir como material didático para pesquisas futuras.

1.7 Estrutura do trabalho

Este trabalho está organizado em 5 capítulos. O capítulo 1 apresenta a introdução do trabalho. No capítulo 2, tem-se a revisão da literatura, abrangendo assuntos como a estrutura dos Colégios Militares; o Programa de Excelência Gerencial (PEG) do Exército Brasileiro, os programas da qualidade e a qualidade no ensino; técnicas estatísticas de AA, ACP e ADI; descoberta de conhecimento em bases de dados.

No terceiro capítulo, é apresentada a metodologia proposta para a elaboração de projetos de exploração de dados na área de ensino, onde no capítulo 4 é apresentada a sua aplicação com dados selecionados e agrupados de diferentes escolas e com múltiplas variáveis.

No capítulo 5, apresenta-se as conclusões obtidas a partir do estudo feito e sugestões recomendadas para trabalhos futuros.

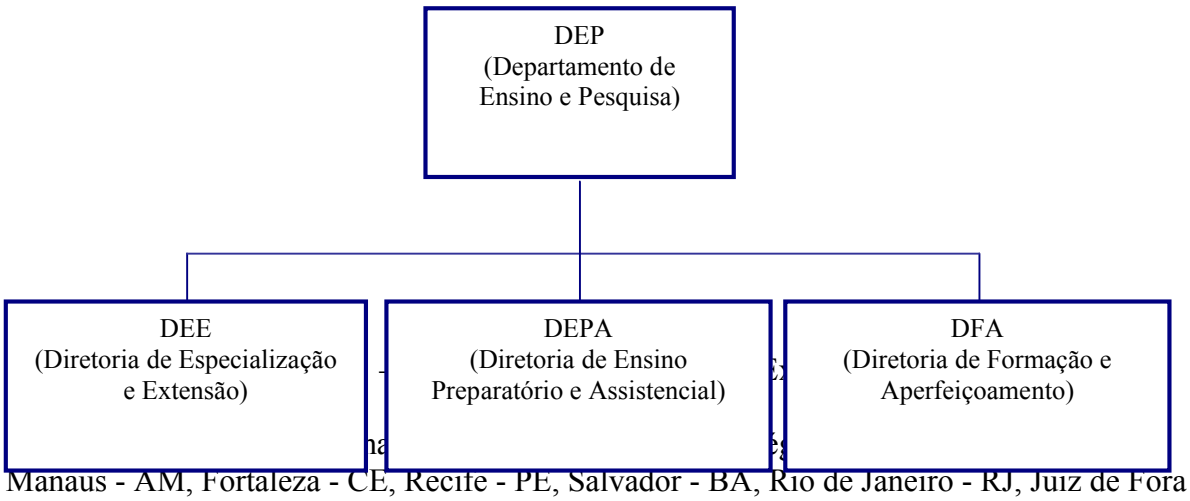
REVISÃO DE LITERATURA

> Precisamos investir para termos retomo e implantar uma cultura de medição em todos os setores. A competitividade só será obtida quando tivermos empresas e governo com sistemas de gestão que garantam qualidade e aprimoramento dos seus processos (Jorge Gerdau Johannpeter, 2003).

Para melhor compreensão, este capítulo está dividido em quatro tópicos, abordando assuntos como a estrutura dos Colégios Militares, a qualidade na gestão do ensino, técnicas estatísticas de AM e descobertas de conhecimento em bases de dados.

1.8 A Estrutura dos Colégios Militares

Segundo as Normas de Planejamento e Conduta do Ensino (NPCE) do Exército Brasileiro de 2005, o Sistema Colégio Militar do Brasil (SCMB) subordina-se diretamente à Diretoria de Ensino Preparatório e Assistencial (DEPA) à qual cabe supervisionar, controlar e coordenar as atividades didático-pedagógicas do Sistema. Conforme a Figura 01, a DEPA é subordinada ao Departamento de Ensino e Pesquisa (DEP), órgão setorial responsável pela condução do Ensino no Exército Brasileiro.



Manaus - AM, Fortaleza - CE, Recife - PE, Salvador - BA, Rio de Janeiro - RJ, Juiz de Fora - MG, Belo Horizonte - MG, Brasília - DF, Campo Grande - MS, Curitiba - PR, Porto Alegre – RS e Santa Maria – RS.

O sistema de ensino, constituído pelos Colégios Militares, é um dos subsistemas do Sistema de Ensino do Exército, e está em consonância com a Legislação Federal, relativa aos ensinos fundamental e médio e com as diretrizes e normas do DEP.

Peculiaridades e características exclusivas identificam-no como o SCMB, conforme consta nos documentos de referência relativos ao Exército, em especial o R-69 – Regulamento dos Colégios Militares, que tem por finalidade estabelecer preceitos aplicáveis a todos os Colégios.

Os Colégios ministram o Ensino Fundamental, da 5ª à 8ª séries, e o Ensino Médio para alunos dependentes de militares e alunos concursados dependentes de civis ou militares em caráter assistencial e preparatório às Escolas Militares e ao Ensino Superior.

Os principais processos são:

a) Ensino – colaborar na formação de cidadãos, intelectualmente preparados e conscientes do seu papel na sociedade, segundo os costumes, valores e as tradições do Exército Brasileiro.

b) Administrativo – apoiar as atividades de ensino e de instrução militar.

Os usuários (Clientes) dos Colégios são os alunos dependentes de militares e alunos concursados dependentes de civis ou militares; pais e responsáveis.

Além deste notado interesse, existe uma variedade de minuciosos processos que, juntos, mantêm a excelência no ensino nacional. Processos esses controlados com rigor, como o Processo de Seleção de Professores, Processo de Elaboração de Provas, Processo Ensino-Aprendizagem, Processo de Controle da Disciplina, Processo de Aquisição de Materiais, todos monitorados por quadros e mapas sumarizados que mostram as principais informações para auxílio na tomada de decisões.

Segundo Falconi (1992), a preocupação atual da alta administração das empresas em todo o mundo tem sido desenvolver sistemas administrativos (*software*) suficientemente fortes e ágeis de tal forma a garantir a sobrevivência das empresas. O Sistema de Gestão Escolar (SGE), *software* utilizado pelos Colégios Militares, controla informações de cadastros, rendimentos, disciplina e saúde dos alunos. Esses dados são centralizados no DEP em uma base de dados única para análise.

1.9 A gestão da qualidade no ensino

Segundo Demo e Ramos (1995), para a escolha da construção de saídas originais para os impasses e desafios do processo educacional, busca-se alguma inovação que contribua de modo significativo para a modificação do atual estado das questões. Propostas visando à recuperação e à aplicação de soluções testadas em outros setores, adaptadas para o setor educacional, vão sendo postas em funcionamento, sendo reconhecidas suas condições como algo específico que se tenta generalizar.

Para Lopes (2004), a globalização de mercados vem provocando traços de instabilidade e incerteza, acompanhando as mudanças, sejam elas profundas ou superficiais, advindas com a implementação de medidas associadas ao neoliberalismo. São formas que se articulam deliberadamente, muitas vezes expressando-se com ceticismo ora na insatisfação, ora na resistência, onde a apatia não pode ser ignorada; são situações de incompatibilidade existencial marcando a sociedade.

Por outro lado, Demo e Ramos (1995) dizem que os anseios por respostas às inquietações provocam a sensação de que, em muitos âmbitos, vivem-se situações precárias que se produzem e até se aprofundam. Há uma demanda latente por melhoria em todos os setores; busca-se a melhoria da qualidade da educação e da situação educacional.

Nas abordagens das questões educacionais, a relação entre educação e qualidade vem ganhando espaço cada vez maior. Recorre-se ao termo qualidade no contexto dos debates sobre educação e como definição do conceito de qualidade em si. Educação recebe tratamento especial, através da maior gama de considerações. Qualidade indica o esforço realizado e o modo como ele pode ser avaliado e valorado frente a outras tentativas.

Apesar das grandes alterações e do progresso fantástico da humanidade nestes últimos séculos, os professores continuam a lecionar da mesma maneira que seus colegas faziam há

cem anos. O sistema educacional encontra-se em plena era industrial, enquanto a sociedade moderna já avança para uma nova etapa: da eletrônica, da informática e da comunicação.

A proposta é defendida como meio de promover mudanças e alcançar como resultado qualidade e excelência. Tal método é visto como solução aos impasses e inquietações oriundos dos processos educacionais e do papel da educação no contexto social. O apelo ao papel da educação como mudança implica o reconhecimento de ser o nodo vigente insatisfatório.

Ramos (1992) afirma que os teóricos da qualidade total conceituam qualidade como atendimento dos interesses, desejos e necessidades do cliente. Para que a qualidade aconteça na escola, são indispensáveis canais de comunicações permanentes com todos os que utilizam os seus serviços, a fim de clarificar o que almejam e satisfazer as expectativas de tais clientes. Também afirma que uma escola que adota a Filosofia da Qualidade trabalha em função do seu cliente maior, o aluno, sendo seu propósito enriquecê-los como ser humano e cidadão. O êxito da escola recai sempre sobre sua capacidade de organizar e promover ações educativas de modo competente e flexível se preciso for, mudando seu estilo de trabalho sempre que sua clientela assim o exigir.

Evidências de qualidade são encontradas em cada etapa do processo de aprendizagem quando há condições suficientes para evitar o fracasso. Os problemas de aprendizagem constatados em avaliações no processo, quando realizados os reforços pedagógicos em oficinas ou atividades de recuperação, evitam as perdas, as repetições desnecessárias, que comprometem o esforço educativo da escola e dos alunos.

Para que se possa garantir a qualidade como processo, é necessário pensar em melhoria constante, pois qualidade, não é algo que se instala, estabelece ou institui de uma única vez. É conquista ou construção ao longo do tempo, através de aperfeiçoamento contínuo. As sugestões para melhorar o trabalho escolar, não surgem apenas das lideranças da escola (formais), podendo ser propostas pelos alunos e pais, agentes decisivos no processo de ensino e aprendizagem.

Ramos (1992) diz que, neste contexto, de melhoria contínua, o "controle" permanente do progresso realizado, o exame dos resultados obtidos e a apresentação de propostas inovadoras conduz a escola a níveis mais elevados da qualidade.

O avanço do conhecimento e de suas aplicações tecnológicas requer que, continuamente, as pessoas tenham oportunidades de aprender novas idéias, novos conceitos e novas habilidades, de rever suas crenças e valores, de ampliar seus horizontes e de ensaiar novas visões de mundo.

A compreensão de que o maior desafio do setor público brasileiro é de natureza gerencial, fez com que, na década de 90 fosse buscado um novo modelo para a gestão pública focado em resultados e orientado para o cidadão. Esse modelo de gestão pública deveria orientar as organizações nessa transformação gerencial e, ao mesmo tempo, permitir avaliações comparativas de desempenho entre organizações públicas brasileiras e estrangeiras, e mesmo com empresas e demais organizações do setor privado.

Segundo o Manual para Avaliação da Gestão Pública do Programa da Qualidade no Serviço Público, em 1997, o Programa da Qualidade no Serviço Público optou pelos Critérios de Excelência utilizados no Brasil, e em diversos países, e que representam o “estado da arte” em gestão. A adoção, sem adaptação dos modelos utilizados pelos prêmios e sistemas existentes, mostra-se inadequada para parte das organizações públicas, principalmente aquelas integrantes da administração direta, em função da natureza dessas organizações e da linguagem, caracteristicamente empresarial, adotada por esses modelos.

A estratégia utilizada pelo Programa foi de adaptação da linguagem, ou seja, de explicação dos conceitos, mantendo o alinhamento com as características essenciais que definem todos os modelos analisados como de excelência em gestão. A adaptação da

linguagem visava interpretar, para o setor público, os conceitos de gestão contidos nos modelos e preservar a natureza pública das organizações que integram o aparelho do Estado brasileiro.

De lá para cá, e sob a mesma orientação, o Modelo de Excelência em Gestão Pública tem passado por aperfeiçoamentos contínuos com o propósito de acompanhar o "estado da arte" da gestão preconizado pelos modelos de referência que lhe deram origem e, também, de acompanhar as mudanças havidas na administração pública brasileira. A Figura 02 representa graficamente o Modelo, destacando a relação entre suas partes.

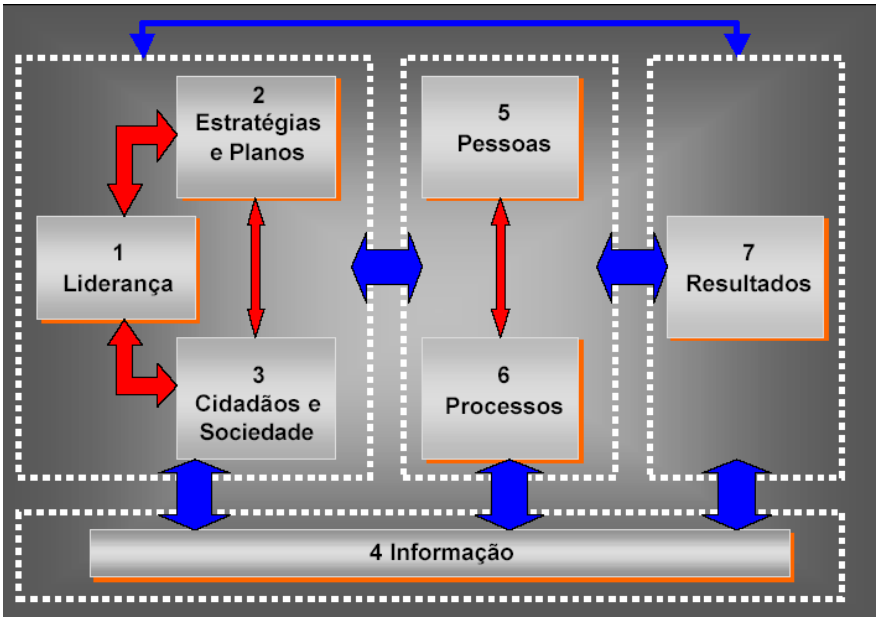


FIGURA 02 – Representação do Modelo de Excelência em Gestão Pública

O Modelo de Excelência em Gestão Pública é a representação de um sistema gerencial constituído de sete partes integradas, que orientam a adoção de práticas de excelência em gestão, com a finalidade de levar as organizações públicas brasileiras a padrões de desempenho e de excelência em gestão.

Liderança, Estratégias e Planos e Cidadãos e Sociedade formam o primeiro bloco e representam o planejamento. Por meio da liderança forte da alta administração, que focaliza as necessidades dos cidadãos destinatários da ação da organização, os serviços/produtos e os processos são planejados para melhor atender esse conjunto de necessidades, levando-se em conta os recursos disponíveis.

O segundo bloco, formado por Pessoas e Processos, representa a execução do planejamento. Nesse espaço, concretiza-se a ação que transforma objetivos e metas em resultados. São as pessoas, capacitadas e motivadas, que operam esses processos e fazem com que cada um deles produza os resultados esperados.

Resultados forma o terceiro bloco e representa o controle. Serve para acompanhar o atendimento à satisfação dos destinatários dos serviços e da ação do Estado, o orçamento e as finanças, a gestão de pessoas, a gestão de fornecedores e das parcerias institucionais, bem como o desempenho dos serviços/produtos e dos processos organizacionais.

O bloco da Informação representa a "inteligência da organização". Nesse bloco, são processados e avaliados os dados e fatos da organização (internos) e aqueles provenientes do ambiente (externos) que não estão sob seu controle direto, mas que, de alguma forma, podem influenciar o seu desempenho. Esse bloco dá à organização a capacidade de agir corretivamente ou para melhorar suas práticas de gestão e, conseqüentemente seu desempenho.

A Figura 02 também apresenta o relacionamento existente entre os blocos (setas maiores) e entre as partes do Modelo (setas menores), evidenciando o enfoque sistêmico do modelo de gestão.

A excelência gerencial, caracterizada pela contínua avaliação, inovação e melhoria da gestão, resulta na otimização de resultados, seja do emprego de recursos, seja dos processos, produtos e serviços. O desempenho das organizações é avaliado constantemente e, a cada ciclo de mensuração, a situação econômica e financeira é retratada.

Gil (1992) defende que a principal missão das entidades é a continuidade operacional, segundo objetivos de atendimento à sociedade à qual elas estão incorporadas. Sobrevivência é a palavra-chave para qualquer tipo de organização.

O indicador de qualidade organizacional tem por finalidade atender à necessidade de quantificação da qualidade a cada momento histórico da entidade. Trabalhar com indicadores de qualidade facilita o processo da qualidade organizacional, permitindo a comparação, via séries históricas, mostrando a evolução das métricas dos indicadores, bem como registra a intensidade da efetividade da ação da qualidade, comparando os indicadores antes e depois de aplicados.

> "A sensibilidade e análise constante no ciclo de vida do indicador de qualidade, a cada estágio vigente, é tarefa crucial do profissional da qualidade organizacional, principalmente, porque a qualidade empresarial é medida por uma família ou cesta de indicadores da qualidade em estágios diferentes de vida" (Gil, 1992).

Indicadores facilitam o planejamento e controle da qualidade, viabilizando a análise comparativa ocorrida em ambientes/linhas de negócios diversificados. Neste trabalho, utilizar-se-á indicadores de rendimento de alunos, como: graus, faltas e comportamento, e, ainda, um indicador social, que é a origem do aluno.

Segundo a metodologia de Gil (1992), os indicadores de qualidade podem perder sua capacidade de retratar a realidade da qualidade organizacional, isto é, podem perder sintonia com ações de qualidade e pontos/situações de revisão da qualidade. Dessa forma, a visão de um ciclo de vida dos indicadores de qualidade se encerra. A qualidade total organizacional é medida e acompanhada segundo seu estágio de vida e, conseqüentemente, capacidade de mensurar os eventos da qualidade organizacional.

Os indicadores construídos segundo os objetivos/interesses de seus consumidores devem representar o espelho da qualidade dos processos e resultados empresariais, atendendo à necessidade de quantificação da qualidade a cada momento histórico da entidade.

1.10 Técnicas estatísticas multivariadas

Os métodos multivariados são apropriados quando as variáveis relacionam-se entre si e estabelecem uma estrutura de dependência, diferentemente da técnica de análise univariada, onde cada variável é considerada individualmente, sem atenção aos inter-relacionamentos.

A AM é a área da análise estatística que se preocupa com as relações entre variáveis dependentes e apresenta duas características principais: os valores das diferentes variáveis devem ser obtidos sobre os mesmos indivíduos, e as variáveis devem ser interdependentes e consideradas simultaneamente. Dessa forma, permite-se um estudo geral das variáveis, pondo em evidência ligações, semelhanças ou diferenças.

A utilização de AM limita-se a sua complexidade de fundamentação teórica e aos recursos computacionais requeridos. Apesar de ter sido desenvolvida há várias décadas, sua

utilização foi dificultada pela complexa manipulação dos dados, bem como dos cálculos efetuados, que, nos dias de hoje, são mantidos através de programas estatísticos específicos.

Atualmente a utilização de AM é aplicada a diversas áreas do conhecimento. Souza, Samohyl e Malavé (2004) utilizaram-se da técnica para realizar uma aplicação em controle de realimentação multivariado, implementando um ajuste de controle proporcional. Lírio (2004) realizou uma aplicação para avaliar a satisfação de clientes de uma empresa de telecomunicações obtendo ótimos resultados, os quais não eram perceptíveis quando as variáveis eram analisadas isoladamente.

Para Lírio (2004), a AM compreende um amplo conjunto de métodos e procedimentos que representam mais de uma característica de uma amostra ou população. Para isso, utiliza-se casos e variáveis em espaços geométricos, escolhem-se métodos, faz-se a máxima economia de hipótese, e transformam-se os dados para visualizá-los num plano ou classificá-los em grupos homogêneos, perdendo o mínimo de informação.

Dados analíticos são normalmente obtidos com objetivo de caracterizar objetos (indivíduos, plantas, animais, solos, climas, amostras de azeite, amostras de sangue, espécies vegetais, espécies animais, doenças, etc...). As avaliações realizadas periodicamente nos Colégios Militares, junto aos indicadores de controle de disciplina, permitem obter mais de dez variáveis para a realização deste estudo.

Esta caracterização é relativamente simples quando o número de variáveis é pequeno. A quantidade de informação que se pode tratar e armazenar sobre um processo é muito grande, complexa e variada. Através da tecnologia dos computadores, na posse de uma enorme quantidade de informações, a questão que surge é como interpretá-las e como extrair informação significante.

Segundo Ferraudo (2005) os seres humanos possuem capacidade única de reconhecer padrões. Toda a aprendizagem está relacionada com a capacidade do cérebro de identificar, isolar, associar e reconhecer formas, sons ou conceitos. Esse processo é de tal forma complexo, que tem apaixonado cientistas e conduzido a tentativas de exploração do seu mecanismo e ao desenvolvimento de metodologias matemáticas, como as redes neurais, ou a inteligência artificial.

A combinação da visão do processo de memorização e reconhecimento confere aos seres humanos habilitações próprias ainda impossíveis de serem executadas por computadores tecnologicamente sofisticados. No entanto, a capacidade humana de identificação por reconhecimento visual vai até a terceira dimensão. A questão fundamental reside em transformar informação m-dimensional em tri ou bidimensional. Isso pode ser feito através de várias técnicas entre as quais se incluem AA, ACP, Escalonamento Multidimensional, Análise de Correspondência, ADI, entre outras.

De acordo com Werkema (1995), as ferramentas estatísticas são utilizadas com o objetivo de reduzir e controlar as incertezas envolvidas na situação de tomada de decisões, além de contribuírem com a redução da variabilidade dos processos analisados. Nesta pesquisa, utilizar-se-á técnicas de AM para redução da dimensionalidade das variáveis que refletem o rendimento escolar em questão e melhor visualização dos resultados pelos administradores.

Segundo Haykin (2001) um problema comum em reconhecimento estatístico de padrões é a seleção das características ou extração de características. A seleção de características se refere a um processo no qual um espaço de dados é transformado em um espaço de características que, em teoria, tem exatamente a mesma dimensão que o espaço original de dados.

Entretanto, a transformação é projetada de tal forma que o conjunto de dados pode ser representado por um número reduzido de características "efetivas" e ainda reter a maioria do

conteúdo de informação intrínseco dos dados, ou seja, o conjunto de dados sofre uma redução de dimensionalidade.

Os métodos utilizados para trabalhar com os dados, neste estudo, são a AA, ACP e ADI. Após utilizar a AA, são empregadas a ACP para auxiliar a conclusão desta análise. Por fim, ainda aplica-se ADI, para classificar alguns alunos.

1.10.1 Análise de Aglomeração

Segundo Ferraudo (2005), os procedimentos de AA surgiram da preocupação inicial de biólogos, antropometristas e psicólogos, em avaliar numericamente as semelhanças ou dissemelhanças entre organismos, com vistas à elaboração de esquemas de classificação.

Tais procedimentos foram rapidamente difundidos em outras áreas de estudo, dada a diversidade de situações em que podem ser utilizados com vantagem, dados seus aspectos numéricos e objetivos. Usualmente, o interesse se volta para o agrupamento de objetos ou indivíduos semelhantes, em termos de suas características (variáveis).

Tal como a Análise Fatorial, a AA estuda todo um conjunto de relações interdependentes. Ela não faz distinção entre variáveis dependentes e independentes. Ao contrário, examina condições de interdependência entre todo o conjunto de variáveis.

Um cluster visa agrupar variáveis com características comuns, sem perder informações de todo o conjunto em estudo. A AA é utilizada nas diversas áreas do conhecimento, por se tratar de uma medida contínua e que possibilita a interpretação individual de cada grupo e a relação que este grupo possui com os demais.

Segundo Regazzi (2001) a AA constitui uma metodologia numérica multivariada com o objetivo de propor uma estrutura classificatória ou do reconhecimento da existência de grupos, objetivando, mais especificamente, dividir o conjunto de observações em um número de grupos homogêneos, segundo algum critério de homogeneidade.

Conforme Malhotra (2001), a maioria dos métodos de aglomeração consiste de processos relativamente simples, não tem um apoio de um raciocínio estatístico rigoroso. Ao contrário, a maioria dos métodos de conglomeração é heurística, baseada em algoritmos. Dessa forma, a AA contrasta significativamente com a análise de variância, a regressão e a ADI, que estão relacionadas com um rigoroso raciocínio estatístico.

Segundo Lírio (2004), a AA, também chamada de análise de conglomerados, é uma técnica usada para classificar casos em grupos homogêneos chamados conglomerados, com base no conjunto de variáveis considerado. Os casos (objetos ou indivíduos) em cada conglomerado tendem a ser semelhantes entre si, mas diferentes de casos em outros conglomerados. A AA tem por finalidade reunir, por algum critério de classificação, as unidades amostrais em grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos.

Tanto a AA quanto a ADI dizem respeito à classificação dos indivíduos em determinado conglomerado. A ADI exige o conhecimento prévio da composição do grupo ou conglomerado para cada objeto ou caso incluídos, para então se definir uma regra de classificação. Na AA não há qualquer informação a priori sobre a composição do grupo ou conglomerado para qualquer de seus objetos. Os grupos ou conglomerados são sugeridos pelos dados, e não definidos, a priori.

Os processos de aglomeração podem se hierárquicos ou não-hierárquicos. Na aglomeração hierárquica é estabelecida uma ordem, ou estrutura em forma de árvore, que produz seqüência de partições em classes cada vez mais vastas. O que não ocorre na

aglomeração não-hierárquica, na qual se produz, diretamente, uma partição em um número fixo de classes.

O método mais comum é o da classificação hierárquica, onde os objetos são agrupados à semelhança de uma classificação taxonômica e representada em um gráfico com uma estrutura em árvore, denominada dendograma.

Embora aconteça perda de informação, esse gráfico é de grande utilidade para a classificação, comparação e discussão de agrupamentos. Esse gráfico apresenta resultados de aglomeração, onde as linhas verticais representam conglomerados unidos e a posição da reta na escala indica as distâncias às quais os conglomerados foram unidos.

O esquema de aglomeração informa sobre objetos ou casos a serem combinados em cada estágio de um processo hierárquico de aglomeração e os centróides de conglomerados representam os valores médios das variáveis para todos os casos ou objetos em um conglomerado particular.

Os centros de conglomerados são os pontos iniciais em um conglomerado não-hierárquico. Os conglomerados são construídos em torno desses centros. A composição de um conglomerado indica o conglomerado ao qual pertence cada objeto ou caso.

A Figura 03 apresenta uma estrutura básica da aplicação da AA, representada em etapas. O primeiro passo consiste em formular o problema de aglomeração definindo as variáveis sobre as quais se baseará a aglomeração.

Logo após, faz-se a coleta dos dados, que serão reunidos numa tabela com ***m*** colunas (variáveis) e ***n*** linhas (objetos).

Antes de escolher a medida de distância para a análise dos dados, é necessário verificar se os mesmos encontram-se com a mesma unidade de medida; caso contrário, deve-se fazer a padronização dos mesmos.

Escolhe-se, então, uma medida apropriada de distância, que irá determinar o quão são semelhantes, ou diferentes os objetos que estão sendo agrupados.

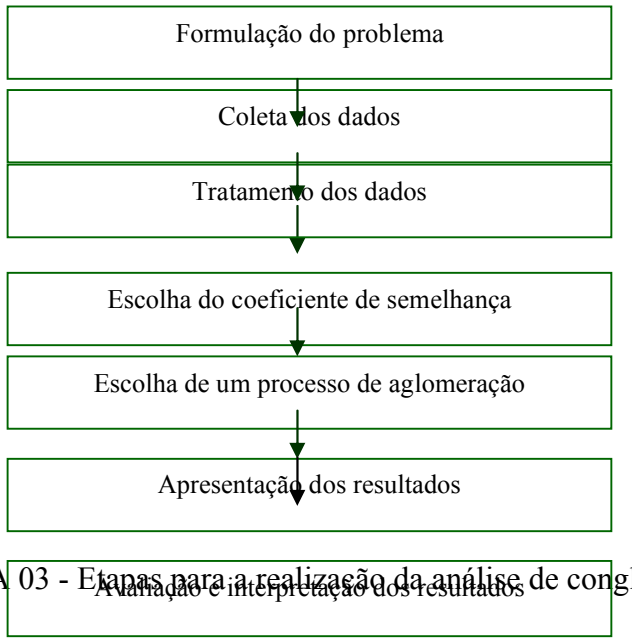


FIGURA 03 - Etapas para a realização da análise de conglomerados.

Dentre vários processos de aglomeração, o pesquisador deve escolher aquele que se parece mais apropriado ao problema estudado.

Com o objetivo de agrupar objetos semelhantes, geralmente, costuma-se avaliar a semelhança em termos de distância entre pares de objetos. Os objetos que possuem a menor distância entre si são mais semelhantes um do outro, do que os objetos com a maior distância. Dentre as várias maneiras de calcular a distância entre dois objetos, está a medida de semelhança mais utilizada, que é a distância euclidiana.

Um grande problema da AA é a escolha da medida de proximidade mais adequada, sendo que as técnicas são baseadas em diferentes medidas de proximidade e nem sempre chegam ao mesmo resultado.

Para proceder esta classificação, faz-se necessário definir matematicamente o que venha ser caracterizado proximidade, ou seja, a distância entre dois objetos, definindo-se, a partir daí, o critério de agrupamento de duas classes. A análise de agrupamentos é constituída de diversas técnicas e algoritmos cujo objetivo é encontrar e separar objetos em grupos similares, utilizando-se um procedimento multidimensional. Trabalha-se com distâncias e técnicas de aglomeração. Entre as medidas mais usuais, para estabelecer o conceito de distância entre dois objetos baseada nos valores de variáveis, podem-se destacar as seguintes formas: distância euclidiana e distância de Mahalanobis.

Considerando o caso mais simples, no qual existem ***n*** indivíduos, onde cada um dos possui valores para ***p*** variáveis, a distância euclidiana entre eles é obtida mediante o teorema de Pitágoras para um espaço multidimensional.

Segundo Manly (1986), a distância euclidiana, quando for estimada a partir das variáveis originais, apresenta a inconveniência de ser influenciada pela escala, pelo número de variáveis e pela correlação existente entre as mesmas. Para contornar as escalas, faz-se a padronização das variáveis em estudo, para que possuam a variância igual à unidade.

Considerando dois indivíduos i e i', a distância entre eles é dada por

$$d_{ii'} = \left[\sum_{j=1}^{p} (X_{ij} - X_{i'j})^2 \right]^{\frac{1}{2}} \quad (1)$$

A medida mais utilizada para a quantificação das distâncias entre duas populações, quando se tem informação sobre a média, variância e covariância residual, e, ainda, existe a repetição de dados, é a distância de Mahalanobis (***D²***). A distância de Mahalanobis considera a variabilidade de cada unidade amostral, sendo recomendada para dados provenientes de delineamento experimentais e, principalmente, quando as variáveis são correlacionadas. Quando as correlações entre as variáveis forem nulas, consideram-se as variáveis padronizadas, a distância de Mahalanobis ***D²*** é equivalente à distância euclidiana.

A forma mais simples de explicar como obter tal medida é a forma matricial, sendo que esta medida entre duas unidades amostrais (tratamentos, indivíduos, populações), i e i', é fornecida por

$$D^2_{ii'} = \left(X_{\sim i} - \overline{X}_{\sim i'}\right) S^{-1} \left(X_{\sim i} - \overline{X}_{\sim i'}\right)' \quad (2)$$

Onde ***S*** é a matriz de dispersão amostral (matriz de correlação) comum a todas as unidades que, no caso de delineamentos experimentais, trata-se da matriz de variâncias e covariâncias.

$X_{\sim i} - \overline{X}_{\sim i'}$ = vetores p-dimensionais de médias i e i', respectivamente, com $i \neq i'$ e i, i' = 1, 2, ..., n. Embora $D^2_{ii'}$, seja o quadrado da distância de Mahalanobis, será chamado de distância de Mahalanobis. A importância dessa medida aumenta a partir do momento em que há repetições, isto é, a unidade amostral consiste num conjunto de indivíduos e, também, quando existe correlação. Esta medida leva em conta a variabilidade dentro das unidades amostrais, e não apenas medidas de tendência central.

Admitindo-se distribuição multinormal p-dimensional e homogeneidade nas matrizes de covariância nas unidades amostrais, pode-se chamar distância generalizada de Mahalanobis. Quanto ao processo de aglomeração, utiliza-se a aglomeração hierárquica, caracterizada pelo estabelecimento de uma estrutura em forma de árvore. O método da variância procura gerar grupos, de modo a minimizar a variância dentro do conglomerado. O processo de Ward é um método de variância bastante utilizado, pois consiste em minimizar o quadrado da distância euclidiana às médias dos aglomerados, ou seja, fundir as duas classes para as quais a perda é menor.

1.10.2 Análise de Componentes Principais

A ACP é um método fatorial cuja característica principal é a redução do número dos caracteres. Esse método não se faz por uma simples seleção de alguns dos fatores, mas pela construção de novos caracteres sintéticos, obtidos pela combinação dos caracteres iniciais, por meio dos fatores, consistindo numa transformação linear das variáveis originais em novas variáveis, de tal forma que a primeira nova variável seja responsável pela maior variação possível, existente no conjunto de dados, de modo análogo à segunda, e demais variáveis, até que toda a variação do conjunto tenha sido explicada.

Assim, a medida da quantidade de informação explicada por cada componente principal é a sua variância. Por essa razão, os componentes principais são ordenados em ordem decrescente da sua variância, ou seja, o componente principal que contém mais informação é o primeiro, sendo o último aquele componente principal com menos informação.

Os componentes principais são não correlacionados, o que é interessante, pois um pesquisador, estando com um problema envolvendo variáveis originais de complexo inter-relacionamento, poderá analisar um conjunto menor de variáveis não correlacionadas, que são os componentes principais.

ACP é considerada uma técnica estatística exploratória, utilizada na tentativa de compreender o inter-relacionamento entre as variáveis originais. A primeira aplicação da ACP foi no campo de testes educacionais.

Em 1901 Karlr Pearson apresentou o método destinado inicialmente ao ajuste de planos em geometria espacial. O objetivo da ACP era o de encontrar linhas e planos que melhor se ajustassem a um conjunto de pontos em um espaço p-dimensional.

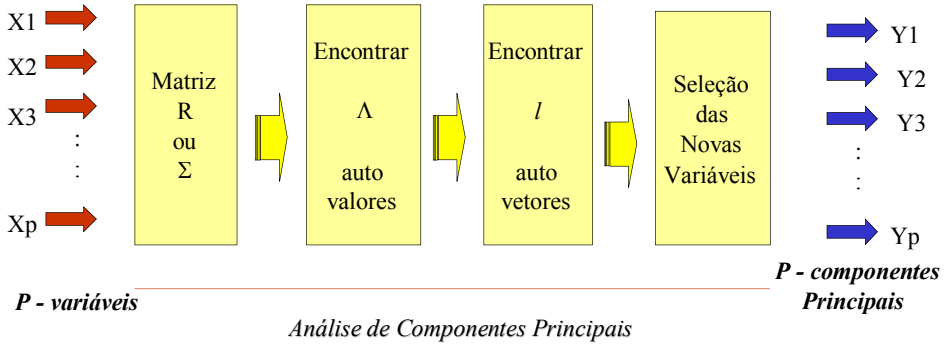
Na década de 30, tornou-se possível sua utilização em análises estatísticas de conjuntos de características para construir um novo conjunto de variáveis, menos numeroso que o original e que resumisse adequadamente a informação contida nas variáveis originais.

Nessa época, um trabalho sobre o desempenho de estudantes foi avaliado por meio de uma seqüência de testes escolares, onde as variáveis utilizadas, na sua maioria, eram correlacionadas. Então, a matriz de correlação e a matriz de covariância foram utilizadas para fazer-se uma análise conjunta. Quando um estudante apresentava boas notas nos testes aplicados, pensava-se que era porque ele possuía algum componente psicológico mais desenvolvido do que os outros, facilitando assim algumas tarefas.

Na Psicologia, as variáveis que apresentavam uma maior influência foram chamadas de fatores mentais. Na Matemática foram denominadas de fatores e, depois, elas receberam o nome de componentes. A componente era determinada pela combinação linear das variáveis que apresentassem a maior variabilidade na matriz de covariância.

A análise que encontrava essas componentes, e que maximizava a variância dos dados originais, foi denominada de *Principal Component Analysis*. Atualmente, um dos principais usos da ACP ocorre quando as variáveis são originárias de processos em que diversas características devam ser observadas ao mesmo tempo.

Primeiramente, é necessário calcular a matriz de variância-covariância (**Σ**), ou a matriz de correlação (**R**), encontrar os autovalores e os autovetores e, por fim, escrever as combinações lineares, que serão as novas variáveis, denominadas de componentes principais (CP). Para o estudo desse item, segue-se o esquema da Figura 04.



Fonte: SOUZA (2000), p 25.

FIGURA 04 - Esquema da aplicação da análise de componentes principais

Segundo Magnusson (2003), estabelecendo-se algumas premissas importantes, e usualmente improváveis, é possível determinar a posição dos eixos no espaço multidimensional, usando-se a álgebra de matrizes.

Para a geração das CP's, suponha-se que $\boldsymbol{X}$ é um vetor de *p*-variáveis aleatórias e que a estrutura de variância e correlação entre as variáveis seja de interesse para estudo.

Se *p* for muito pequeno ou se as correlações entre as variáveis forem muito pequenas, a investigação das variáveis individualmente deve ser preferida. Caso isso não ocorra, pode-se utilizar a metodologia de ACP, que possibilita investigar poucas CP's, ao invés de todo o conjunto das variáveis originais, mantendo-se a maioria das informações das matrizes de variância e correlação.

Considera-se o vetor aleatório $\boldsymbol{X}' = [X_1, X_2, ..., X_P]$, do qual calcula-se a matriz de variância-covariância $\boldsymbol{\Sigma}$ e média $\boldsymbol{\mu_{xi}}$, quando se considera o caso populacional. Neste estudo, utiliza-se apenas um conjunto amostral; logo, a matriz $\boldsymbol{\Sigma}$ será estimada através da matriz de variância-covariância amostral $\boldsymbol{S}$ e vetor média $\overline{\overline{X}} = [\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p]$.

A partir da matriz $\boldsymbol{S}$ é possível encontrar os valores $\hat{\Lambda}_1 \geq \hat{\Lambda}_2 \geq ... \geq \hat{\Lambda}_p \geq 0$, que são as raízes características, todas distintas e apresentadas em ordem decrescente de valores e, como $\boldsymbol{S}$ é positiva definida, todos os autovalores são não negativos.

Ao se estudar um conjunto de *n* observações de *p-variáveis*, é possível encontrar-se novas variáveis denominadas de $\hat{Y}_k$, $k = 1, ..., p$, que são combinações lineares das variáveis originais X_p, não-correlacionadas e apresentam um grau de variabilidade diferente umas das outras.

A primeira componente, extraída da matriz de dados originais ***X,*** é uma combinação linear representada por

$$\hat{Y}_1 = \hat{\ell}_{11} X_1 + \cdots + \hat{\ell}_{p1} X_p = \hat{\ell}'_1 X \tag{3}$$

cuja variância amostral dada por:

$$S^2_{Y_1} = \sum_{i=1}^{p} \sum_{j=1}^{p} \hat{\ell}_{i1} \hat{\ell}_{j1} S_{ij} = \hat{\ell}'_1 S \hat{\ell}_1 \tag{4}$$

é a maior dentre as possíveis combinações lineares de $X_1, X_2, ..., X_P$, sob a restrição de que $\hat{\ell}'_1 \hat{\ell}_1 = 1$.

Segundo Morrison (1976), para se determinar os coeficientes, introduz-se a restrição de normalização por meio do multiplicador de Lagrange $\hat{\Lambda}_1$ e diferencia-se em relação a $\hat{\ell}_1$, uma vez que o objetivo é maximizar a variância, sujeita à restrição $\hat{\ell}'_1 \hat{\ell}_1 = 1$. A primeira derivada da função de Lagrange, em relação a $\hat{\ell}_1$ está representada por

$$\frac{\partial}{\partial \hat{\ell}_1} \left[S^2_{Y_1} + \hat{\Lambda}_1 \left(1 - \hat{\ell}'_1 \hat{\ell}_1 \right) \right] = \frac{\partial}{\partial \ell_1} \left[\hat{\ell}_1 S \hat{\ell}_1 + \hat{\Lambda}_1 \left(1 - \hat{\ell}'_1 \hat{\ell}_1 \right) \right] = 2 \left(S^2_{Y_1} - \hat{\Lambda}_1 I \right) \hat{\ell}_1 \tag{5}$$

onde os coeficientes encontrados devem satisfazer as *p-equações* lineares simultaneamente.

Ao solucionar esta equação

$$(S - \hat{\Lambda}_1 I) \hat{\ell}_1 = 0 \tag{6}$$

o valor de $\hat{\Lambda}_1$ deve ser escolhido de modo que

$$\left| S - \hat{\Lambda}_1 I \right| = 0 \tag{7}$$

onde $\hat{\Lambda}_1$ é a maior raiz característica ou autovalor da matriz ***S*** e $\hat{\ell}_1$ é o seu autovetor associado. Para determinar quais das *p-raízes* devem ser utilizadas, pré-multiplica-se a equação (6) por $\hat{\ell}'_1$. Desde que $\hat{\ell}'_1 \hat{\ell}_1 = 1$, obtém-se:

$$\hat{\Lambda}_1 = \hat{\ell}'_1 S \hat{\ell}_1 = S^2_{Y_1} \tag{8}$$

Segundo Morrison (1976), os autovetores associados com o maior autovalor $\hat{\Lambda}_1$ da matriz amostral ***S*** são únicos, pois eles são escalonados de modo que $\hat{\ell}'_1 \hat{\ell}_1 = 1$; a raiz característica $\hat{\Lambda}_1$ é interpretada como a variância amostral de $\hat{Y}_1$.

$$\hat{Y}_2 = \hat{\ell}_{12} X_1 + \cdots + \hat{\ell}_{p2} X_p = \hat{\ell}'_2 X \tag{9}$$

A segunda CP, de maneira análoga, é representada por

$$\hat{\ell}'_2 \hat{\ell}_2 = 1 \tag{10}$$

$$\hat{\ell}_1'\hat{\ell}_2 = 0$$

de modo que a variância de $\hat{Y}_2$ seja máxima. A primeira restrição é feita para que o sistema tenha solução única, e a segunda requer que $\hat{\ell}_1$ e $\hat{\ell}_2$ sejam ortogonais. A conseqüência imediata da ortogonalidade é que as CP's são independentes.

Os coeficientes da segunda componente são encontrados introduzindo-se as restrições através dos multiplicadores de Lagrange $\hat{\Lambda}_2$ e μ diferenciando em relação a $\hat{\ell}_2$.

Se o lado direito da equação

$$\frac{\partial}{\partial \hat{\ell}_2}\left[\hat{\ell}'_2 S\hat{\ell}_2 + \hat{\Lambda}_2\left(1-\hat{\ell}'_2\hat{\ell}_2\right)+\mu\hat{\ell}'_1\hat{\ell}_2\right] == 2\left(S-\hat{\ell}_2 I\right)\hat{\ell}_2 + \mu\hat{\ell}_1 \tag{11}$$

for igualado a zero e pré-multiplicado por $\hat{\ell}'_1$, obtém-se das condições de normalização e ortogonalidade que

$$2\hat{\ell}'_1 S\hat{\ell}_2 + \mu = 0 \tag{12}$$

O segundo vetor deve satisfazer

$$(S-\hat{\Lambda}_2 I)\hat{\ell}_2 = 0 \tag{13}$$

e resulta que os coeficientes da segunda componente são então os elementos do vetor característico correspondendo à segunda maior raiz característica. As restantes CP's são encontradas através dos outros vetores característicos.

Segundo a definição apresentada por Morrison (1976), a *j-ésima* componente principal de uma amostra de *p-variáveis* é uma combinação linear, tal que:

$$\hat{Y}_j = \hat{\ell}_{1j}X_1 + \cdots + \hat{\ell}_{pj}X_p \tag{14}$$

Seus coeficientes são os elementos do vetor característico da amostra da matriz de variância-covariância ***S***, correspondendo à *j-ésima* maior raiz característica $\hat{\Lambda}_j$. Se $\hat{\Lambda}_i \neq \hat{\Lambda}_j$, os coeficientes da *i-ésima* e *j-ésima* componentes são necessariamente ortogonais; se $\hat{\Lambda}_i = \hat{\Lambda}_j$, os elementos podem ser escolhidos para serem ortogonais, mesmo existindo uma infinidade desses vetores. A variância amostral da *j-ésima* componente é $\hat{\Lambda}_j$, e a variância total do sistema é

$$\hat{\Lambda}_1 + \cdots + \hat{\Lambda}_p = tr\,S \tag{15}$$

O grau de explicação fornecido pela *j-ésima* componente é fornecida por

$$\frac{\hat{\Lambda}_j}{tr\,S} \tag{16}$$

Na interpretação da CP, o sinal algébrico e a magnitude de $\hat{\ell}_{ij}$ indicam a direção e a importância da contribuição da *i-ésima* resposta para a *j-ésima* componente, e a variância generalizada S_{yy} representa a dispersão total dos dados, sendo encontrada somando-se as variâncias das variáveis, conforme mostra a relação

$$S_1^2 + S_2^2 + \cdots + S_p^2 = \hat{\Lambda}_1 + \hat{\Lambda}_2 + \cdots + \hat{\Lambda}_p = tr\, S_{yy} \tag{17}$$

Até o momento, derivaram-se as CP's utilizando-se a matriz de variância ***S***, mas a utilização desta matriz para a geração das CP's leva em consideração as unidades amostrais das variáveis envolvidas no processo e a magnitude destas variáveis.

Procurando-se eliminar a influência que uma variável possa causar sobre a outra na formação da componente, utilizar-se-ão as CP's derivadas de variáveis padronizadas. A padronização é feita por meio da relação $Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$, onde a representação matricial será igual a $\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu})$, onde $E(\mathbf{Z}) = \mathbf{0}$ e a covariância será $\mathbf{Cov(Z)} = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \mathbf{R}$.

Dessa forma, utilizando os dados padronizados, garante-se que todas as variáveis tenham o mesmo grau de importância; portanto, trabalha-se com o conjunto de dados padronizados.

Nesse caso, faz-se necessário estimar a matriz ***R*** para se calcularem os autovalores e autovetores que darão origem às componentes principais, cujo procedimento para a estimação dos autovalores e autovetores será o mesmo mostrado anteriormente, apenas substituindo ***S*** por ***R***. Os autovetores passarão a ser denominados de $\hat{e}_p$, pois essa nova representação indica que o conjunto amostral dos dados foi padronizado.

Logo, os pares de autovalores e autovetores estimados da amostra analisada serão representados por $(\hat{\Lambda}_1, \hat{e}_1)$, $(\hat{\Lambda}_2, \hat{e}_2)$, ... ,$(\hat{\Lambda}_p, \hat{e}_p)$; onde $\hat{\Lambda}_1 \geq \hat{\Lambda}_2 \geq \ldots \geq \hat{\Lambda}_p \geq 0$; e fornecerão as novas combinações lineares, expressas por: $\hat{Y}_1 = \hat{e}_1' Z$, $\hat{Y}_2 = \hat{e}_2' Z$, ... , $\hat{Y}_p = \hat{e}_p' Z$.

Segundo Jackson (1981), as combinações lineares obtidas através das CP's, possuem a característica de que nenhuma combinação linear das variáveis originais irá explicar mais que a primeira componente e, sempre que se trabalhar com a matriz de correlação, as variáveis não sofrerão influência das magnitude de suas unidades medidas.

Em controle de qualidade, a ACP é utilizada para reduzir o número de variáveis analisadas pois a monitoração do processo feita com as primeiras componentes mantém um bom grau de explicação das variáveis originais.

A definição do número de componentes a serem utilizadas é feita por meio de dois critérios:

- O primeiro, denominado de método gráfico, representa graficamente a porcentagem de variação explicada pela componente nas ordenadas e, os autovalores em ordem decrescente nas abscissas. Esse critério considera as componentes anteriores ao ponto de inflexão da curva.

- O segundo critério de seleção consiste em incluir somente aquelas componentes cujos valores próprios sejam superiores a 1. Esse critério tende a incluir poucas componentes quando o número de variáveis originais é inferior a vinte e, em geral, utilizam-se aquelas componentes que conseguem sintetizar uma variância acumulada em torno de 70%.

A determinação da variável que possui maior influência na combinação linear será encontrada através da correlação mostrada a seguir

$$r_{\hat{Y}_i, X_k} = \frac{\hat{e}_{ki}\sqrt{\hat{\Lambda}_i}}{\sqrt{s_{kk}}} \; ; \quad i,k = 1, 2, ..., p \tag{18}$$

$$r_{\hat{Y}_i, Z_k} = \hat{e}_{ki}\sqrt{\hat{\Lambda}_i} \; ; \; i,k = 1, 2, ..., p \tag{19}$$

que indica, através de seus valores absolutos, as variáveis que exercem maior influência sobre a componente principal. Segundo Johnson e Wichern (1992), o sinal da correlação indica o modo dessa influência , sendo assim identificada aquela que deve ser monitorada para manter o sistema estável.

A equação (18) deve ser utilizada quando os autovetores forem derivados da matriz de variância ***S***, e a equação (19) quando os autovetores forem derivados da matriz de correlação ***R***.

Segundo Johnson e Wichern (1992), uma ACP freqüentemente revela características que não foram previamente consideradas, e assim, permite interpretações que não iriam, de outro modo, aparecer. Num conjunto grande de variáveis nem todas têm quantidade de informação relevante podendo através da ACP selecionar aquelas que mais possuem esta característica.

Embora, do ponto de vista matemático, a Análise de Agrupamentos apresente uma metodologia bastante simples, e a ACP uma metodologia mais complexa, espera-se uma boa concordância entre os resultados de ambas, devendo uma ser utilizada como complemento da outra.

Na ACP as variáveis não são discriminadas como independentes ou dependentes como na análise de regressão. Todas são tratadas como variáveis. A técnica pode ser entendida como um método de transformação das variáveis originais em novas variáveis não correlacionadas.

1.10.3 Análise Discriminante

Segundo Virgillito (2004), a ADI encontra aplicação nos casos de pesquisas em que, dado um certo número de grupos, haja a necessidade de identificar a qual dos grupos pertence um certo caso (indivíduo ou observação).

A ADI é a técnica de dependência multivariada mais utilizada, É aplicada quando a variável dependente é qualitativa e possui duas ou mais características e as variáveis independentes são quantitativas.

Segundo Malhotra (2001), os objetivos da ADI são:

a) Estabelecer funções discriminantes**,** ou combinações lineares das variáveis independentes ou prognosticadoras, que melhor discriminem entre as categorias da variável dependente (grupos);

b) Verificar se existem diferenças significativas entre os grupos, em termos das variáveis prognosticadoras;

c) Determinar as variáveis preditoras que mais contribuam para as diferenças entre grupos;

d) Enquadrar, ou classificar os casos, em um dos grupos, com base nos valores das variáveis preditoras;

e) Avaliar a precisão da classificação.

As técnicas de ADI são definidas pelo número de categorias que a variável dependente possui. Quando a variável dependente tem duas categorias, a técnica é conhecida como ADI de dois grupos. Quando estão em jogo três, ou mais categorias, temos a ADI múltipla.

A distinção principal é que, no caso de dois grupos, é possível deduzir apenas uma função discriminante, e na ADI múltipla, pode-se calcular mais de uma função.

Após a definição dos grupos, são coletados dados individuais dos elementos de cada grupo. A ADI procura estimar a combinação linear das características individuais, de cada elemento, que melhor discrimina entre os grupos pré-estabelecidos.

A função discriminante utiliza a regressão para processar as variáveis. O objetivo da regressão passo a passo é selecionar, em um grande número de variáveis prognosticadoras, um pequeno subconjunto de variáveis que respondam pela maior parte da variação na variável dependente. Nesse processo, as variáveis prognosticadoras entram na equação de regressão, ou saem, dela, uma de cada vez.

Segundo Malhotra (2001), há várias abordagens para a regressão passo a passo. Na Inclusão Avançada, inicialmente, não há variáveis prognosticadoras na equação de regressão. São introduzidas uma de cada vez, somente se satisfizerem certos critérios definidos em termos da razão F. A ordem em que as variáveis são incluídas baseia-se na contribuição para a variância explicada.

Na Eliminação para Trás, inicialmente, todas as variáveis prognosticadoras são incluídas na equação de regressão. Remove-se, então, as prognosticadoras, uma de cada vez, com base na razão F.

Na solução passo a passo, combina-se a inclusão antecipada com a remoção de prognosticadoras que não mais satisfaçam o critério especificado em cada passo.

Os processos passo a passo não resultam em equações ótimas de regressão, no sentido de gerar o maior R^2 para um número determinado de prognosticadores. Em razão das correlações entre prognosticadores, pode ocorrer que uma variável importante nunca venha a ser incluída, enquanto que variáveis menos importantes possam ser introduzidas na equação.

Para identificar uma equação ótima de regressão, ter-se-ia que calcular soluções combinatórias em que se examine todas as combinações possíveis. Não obstante, a regressão passo a passo pode ser útil quando o tamanho da amostra é grande em relação ao número de prognosticadoras.

Segundo Ferraudo (2005), outra vantagem da ADI é reduzir o espaço dimensional das variáveis independentes para um número de grupos estabelecidos *a priori*. No caso de dois grupos apenas, a análise é transformada em uma única dimensão. É utilizada também para classificar novos elementos dentro de um dos grupos. A ADI implica em obter um valor teórico que é uma combinação linear das variáveis independentes que discrimine melhor entre os grupos definidos *a priori* segundo:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + W_3X_{3k} + ...+ W_nX_{nk} \quad (20)$$

onde:
Z_{jk} = valor da função discriminante j para o objeto k;
a = constante;
W_i = ponderação discriminante para a variável independente i.
X_{ik} = variável independente i para o objeto k.

Supõe-se que as variáveis independentes venham de amostras de populações com distribuição normal multivariada e que se tenha, nos grupos, homogeneidade nas matrizes de variância/covariância das variáveis.

Para a determinação do ponto de corte, se os grupos forem de mesmo tamanho, o ponto ótimo está na metade do caminho entre os centróides dos dois grupos assim definido:

$$Z_{CE} = \frac{Z_A + Z_B}{2} \tag{21}$$

Onde:
Z_{CE} = ponto ótimo de corte para grupos de mesmo tamanho;
Z_A = centróide do grupo A;
Z_B = centróide do grupo B;

Para a determinação do ponto de corte para grupos de tamanhos diferentes, é feita uma média ponderada dos centróides proporcionalmente aos tamanhos de cada grupo, assim definido:

$$Z_{CD} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} \tag{22}$$

Onde:
Z_{CD} = ponto ótimo de corte para grupos de tamanhos distintos;
Z_A = centróide do grupo A;
Z_B = centróide do grupo B;
N_A = número de elementos do grupo A;
N_B = número de elementos do grupo B.

Quanto ao critério de classificação para validar a função discriminante, devem-se obter amostras aleatórias, criando-se dois grupos. Um grupo é utilizado para a obtenção da função discriminante, e o outro para validar a função, criando a matriz de classificação. O critério de classificação de cada objeto no grupo é assim definido:

Classificar um indivíduo dentro do grupo A se $Z_n < Z_{CT}$
Classificar um indivíduo dentro do grupo B se $Z_n > Z_{CT}$

Onde:
Z_n = pontuação Z discriminante para o n-ésimo indivíduo;
Z_{CT} = valor da ponto de corte ótimo.

A Função Discriminante Linear transforma a observação multivariada X, de dimensão p, na observação univariada Y (escore), tal que os escores obtidos para as populações Φ_1, e de Φ_2, sejam separados ao máximo. Sendo μ_1 e μ_2 e Σ, respectivamente, os vetores médios de Φ_1 e de Φ_2, e a matriz de covariância comum a ambas as populações, tem-se a função a seguir.

$$y = (\mu_1 - \mu_2)'\Sigma^{-1} X \tag{23}$$

Então, pode-se expressar a regra de classificação para X_0, como:

Alocar X_0 em Φ_1 se y_0 - m >= 0 ou alocar X_0 em Φ_2 se y_0 - m < 0.

Na realidade, os parâmetros μ_1 e μ_2 e Σ não são conhecidos. Assim, trabalha-se com os seus estimadores: X_1, X_2 e S_p, obtidos de amostras aleatórias dos grupos G_1 e G_2 com tamanhos n_1 e n_2, respectivamente.

A Função Discriminante Linear de Fisher é dada pela expressão:

$$y = (X_1 - X_2)' S_p^{-1} X \tag{25}$$

O valor de corte m é estimado por:

$$m = \frac{1}{2}(y_1 + y_2) \tag{26}$$

onde Y_1 e Y_2 são as médias dos escores para G_1 e G_2. A regra de classificação fica:

Alocar X_0 em G_1 se

$$(X_1 - X_2)' S_p^{-1} X \quad >= \text{ m}, \tag{27}$$

ou alocar X_0 em G_2 se

$$(X_1 - X_2)' S_p^{-1} X < \text{ m} \tag{28}$$

1.11 Descoberta de conhecimento em bases de dados

Segundo Drucker (2000), a Era do Conhecimento está emergindo e, diferentemente da Era Industrial, nesta nova sociedade, a criação e o gerenciamento do conhecimento serão fatores decisivos no ambiente competitivo. Nenhuma empresa conseguirá se manter competitiva se não houver ferramentas de apoio nas áreas de produção, de estatística, de planejamento e de qualidade.

Para o aprendizado tomar forma, com o uso de ferramentas adequadas, dados de diferentes fontes devem ser agrupados e organizados de maneira consistente e proveitosa. *Data wherehousing* é o conjunto refinado de dados armazenados que permite que as empresas memorizem o que ocorre com seus clientes. *Data warehousing* caracteriza o empreendimento com a memória, a qual não tem utilidade sem a ação da inteligência. DM age nos dados com a inteligência. São ações com inteligência sobre as bases de dados organizadas que leva a rever padrões, alterar regras, gerar novas idéias e predizer o futuro.

> ″Informação sozinha não é poder. Se fosse, os bibliotecários dominariam o mundo. Conhecimento sozinho não é poder. Se fosse, os Ph.D.s receberiam os mais altos salários da Terra. Ação Inteligente é poder. Mudança é poder. Além do mais, Informação não é uma coisa para se ter, é, algo para se encontrar. Culto não é o que sabe tudo. É o que sabe onde encontrar tudo na hora certa para transformar Dificuldades em Ação Inteligente″ (Maurício Góis, 2001).

DM é um esforço cooperativo entre as pessoas e os computadores. Os seres humanos planejam o banco de dados. Os computadores pesquisam dados com agilidade e segurança, utilizando padrões matemáticos em conjuntos de dados.

Trata-se de um campo interdisciplinar que reúne conceitos e técnicas de ciência da computação, estatística, inteligência artificial, visualização (design), desenvolvimento de hardware e gestão do conhecimento. A interdisciplinaridade, de um lado, enriquece o tema, mas, de outro, faz com que ele represente coisas diferentes para diferentes pesquisadores.

A Figura 05 apresenta as etapas para um projeto de descoberta de conhecimento em bases de dados.

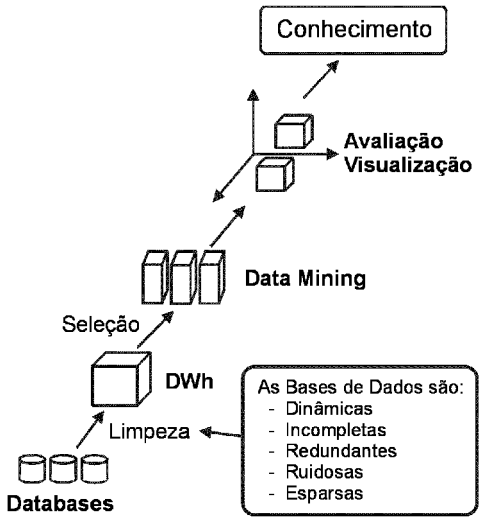


FIGURA 05 – Etapas da descoberta de conhecimento em bases de dados.

Segundo Stair (1998), com o crescimento das empresas aumenta a necessidade de controle e processamento de informações, que, futuramente, poderá não suprir as necessidades de agilidade e produtividade requeridas pelo mercado.

Conforme Silberschatz (1999), a disponibilidade de dados on-line tem crescido, e os administradores têm explorado sua fonte de dados para melhorar a tomada de decisões. Pode-se adquirir muitas informações usando pesquisas simples, pois, utilizando-se grandes bases de dados, não se consegue acessar estes dados através de uma planilha eletrônica.

Cornesky (1993) complementa afirmando que, se não usamos bases de dados coerentes para resolver problemas, não existe razão para armazená-los. Resultados de bancos de dados, bem implementados, podem apresentar sinais para realizar mudanças.

A Figura 06 apresenta os tipos de informações e dados gerados nas empresas, com seu graus de utilidade.

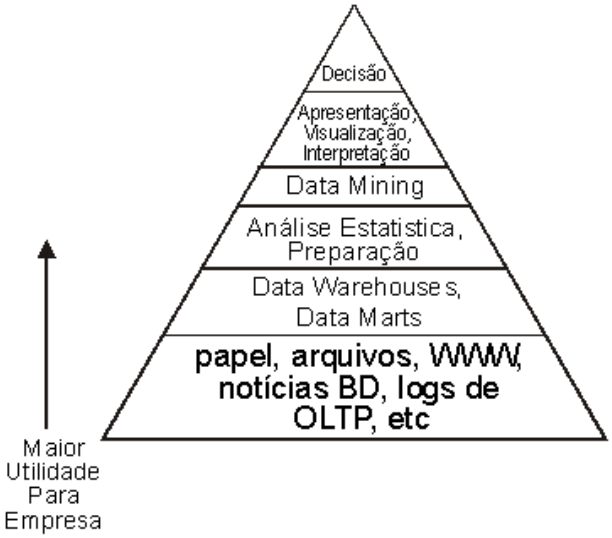


FIGURA 06 – Diagrama dos tipos de informações e dados gerados nas empresas.

Pereira e Fonseca (1997) apresentam o seguinte conceito de Sistemas de Informações: "são mecanismos de apoio à gestão, desenvolvidos com base na tecnologia de informação e com o suporte da informática, para atuar como condutores das informações que visam facilitar, agilizar e otimizar o processo decisório nas organizações."

O armazenamento de dados ficou mais fácil com a grande capacidade computacional a baixo custo, isto é, o custo da capacidade de processamento e armazenamento sofreu sensível queda. Também houve a introdução de novos métodos de aprendizado de máquina para representação do conhecimento baseado em programação lógica, além da análise estatística de dados. Os novos métodos tendem a ser usados intensivamente de acordo com a demanda de processamento e necessidade de extração de informações diversas.

Tendo concentrado muita atenção no armazenamento de dados, o problema passou a ser o que fazer com estes recursos valiosos. A informação está no centro de operações administrativas e as pessoas poderiam fazer melhor uso delas.

Segundo Dachs (1998), o papel do computador, como ferramenta de trabalho em pesquisa e estatística, tem crescido muito. Como ferramenta para investigação heurística, os sistemas de informática estão cada vez mais presentes no processo de busca de novas técnicas e metodologias. O número de pacotes de programas de computador para análises estatísticas está cada vez mais presente nas instituições modernas.

A escolha da combinação de técnicas, para serem aplicadas numa particular situação, depende da natureza das tarefas de pesquisa de DM e da natureza dos dados avaliados. Classificação, estimação, predição, agrupamento por afinidade, clusterização e descrição são algumas das tarefas que caracterizam uma exploração de dados.

Algumas dessas tarefas identificam-se com mineração *top-down*, chamada teste de hipóteses. Nesse tipo de exploração, o comportamento dos dados armazenados no banco de dados é utilizado para aprovar, ou reprovar, idéias e descobrir afinidade relativa dos dados.

Outras tarefas são apropriadas à mineração *bottom-up,* chamada descoberta do conhecimento. Nesta não se criam hipóteses iniciais na exploração. Os dados são induzidos a falarem por si só. Esse tipo de exploração subdivide-se em duas áreas: Direto e o Indireto.

No direto, o estudo procura explicar ou categorizar alguns campos de dados de acordo com a forma como eles rendem ou respondem.

Melhorar o ensino significa qualificar o seu produto. É necessário mensurar estatisticamente as variáveis que representam os fatores de qualidade de ensino. A intenção é

descobrir novas informações com a intensa aplicação de diferentes variáveis de pesquisa, o que determina um desafio novo e incerto, características da aplicação de DM.

Os Sistemas de Gerenciamento de Banco de Dados deram segurança e performance para o acesso e armazenamento de informações, mas essa é só uma pequena parte do que poderia ser obtido dos dados. Sistemas tradicionais de processamento de transações *on-line* são ferramentas boas no que se refere a repor dados rapidamente e com segurança em bancos de dados, mas não são boas para retornar uma análise significativa relativa aos dados.

Dados analisados podem prover conhecimento adicional sobre a área de ensino e derivar novos conhecimentos sobre os desempenhos escolares. DM ou KDD têm benefícios óbvios para esse estudo, pois permitem as extrações não triviais, previamente desconhecidas, de informações potencialmente úteis dos dados. Isto engloba várias técnicas de aproximações diferentes, como *clustering*, sumarização de dados, regras de aprendizado de classificação, análise de mudanças e detecção de anomalias.

DM refere-se ao "uso de uma variedade de técnicas para identificar informações úteis em bancos de dados e a extração dessas informações, de tal maneira que possam ser usadas em áreas tais, como: teoria da decisão, estimação, predição e previsão. Os bancos de dados são geralmente volumosos e, na forma como se encontram, nenhum uso direto pode ser feito deles; as informações escondidas nos dados é que são realmente úteis" (*Clementine User Guide*, 2000).

DM relaciona-se com a análise de dados e o uso de ferramentas computacionais (*softwares*) na busca de características, regras e regularidades em um grande conjunto de dados. DM é um esforço cooperativo entre os seres humanos e os computadores. Os seres humanos planejam o banco de dados; os computadores pesquisam através de dados, procurando por padrões que correspondam às metas.

Utilizar-se-á esses produtos e técnicas para a realização de um estudo que crie uma ferramenta para acompanhamento da dinâmica do desempenho escolar. Precisar-se-á trabalhar em cima do registro dos eventos qualitativos. Para isso são fixadas premissas de natureza ontológica e semântica, bem como a análise de seu comportamento e suas relações com outros eventos.

Há três razões principais para se desenvolver um projeto de exploração de dados:

- Visualização dos dados: as instituições precisam dar significado a uma quantidade cada vez maior de informações em seus bancos de dados. Antes de realizar qualquer análise, o objetivo é qualificar e armazenar os dados a serem trabalhados e encontrar novas formas de visualizá-los de forma mais natural e transparente para os usuários.

- Descoberta de novos conhecimentos**:** a maior parte das aplicações atuais de DM se enquadra nesta tecnologia, cujo objetivo é explicitar relacionamentos ocultos, padrões e correlações entre os diferentes dados existentes no banco de dados da empresa.

- Acuracidade dos dados: muitas vezes, as empresas descobrem que seus dados são incompletos, errados ou contraditórios. Decorre disso a necessidade de se obterem dados cada vez mais consistentes para processamento e análise futuros.

O armazenamento e a recuperação de dados, para suporte à decisão, conduzem a vários pontos importantes. Embora muitas consultas possam ser escritas em SQL, outras não são expressas, ou não são facilmente expressas.

Linguagens de consulta a banco de dados não são apropriadas para a análise estatística de dados. Há vários programas que auxiliam na análise estatística. Esses programas possuem interfaces para bancos de dados, possibilitando a armazenagem e recuperação das informações para análise.

Técnicas para descoberta de conhecimento, desenvolvidas pela comunidade de inteligência artificial, tentam encontrar regras e modelos estatísticos a partir dos dados. O campo de DM (extração de dados) combina idéias de descoberta de conhecimento, com a implementação eficiente de técnicas, que possibilitem usá-las em banco de dados muito grandes.

Para pesquisas eficientes em dados tão diversos, as instituições passaram a construir *data warehouses* (depósitos de dados). Os *data warehouses* colecionam informações de diversas fontes sob um esquema unificado em um único local.

Embora seja preferível deixar as análises estatísticas complexas para os programas estatísticos, os banco de dados podem aceitar formas de análises de dados simples e utilizadas com maior freqüência. Uma vez que, em geral, há grande volume de dados em um banco de dados, precisam ser resumidos de alguma forma para derivar informações que o usuário possa utilizar. Geralmente se utilizam as funções agregadas para esse fim.

Como a agregação de funcionalidade à SQL é limitada, várias aplicações foram implementadas por diferentes bancos de dados. Embora a SQL defina apenas algumas funções agregadas, muitos sistemas de bancos de dados oferecem um conjunto mais rico de funções, incluindo variância, mediana e outras. Alguns sistemas permitem adicionar novas funções agregadas.

Segundo Louzada Neto (2002), com a disponibilidade de grandes recursos computacionais, a baixo custo, é extremamente fácil acumular informação. Várias organizações geram e coletam grandes volumes de dados de suas operações diárias. Nesse contexto, surgem várias questões.

Tendo acumulado uma grande quantidade de dados, o que fazer com eles? Como reverter essas informações em benefícios para a própria organização? Como os pesquisadores ou executivos podem identificar e utilizar as informações escondidas nos dados coletados fazendo com que estas informações se revertam em vantagens em um tempo rápido para a organização?

Nesse contexto, surge o termo DM, ou mineração de dados, o qual é um dos vários termos utilizados para descrever o conceito de busca de conhecimentos em bancos de dados. DM consiste no processo de extrair, sem conhecimento prévio e essencialmente inteligível, informação de um grande banco de dados e de utilizar esta informação para a tomada de decisões.

Várias empresas bem sucedidas estão optando por DM no auxílio à tomada de melhores decisões. A partir do uso de técnicas analíticas eficientes, o DM possibilita a transformação dos dados em informações importantes ao desenvolvimento de estratégias para aumentar a produtividade e qualidade de seus serviços. O DM faz parte de um ciclo contínuo - um processo que combina dados acumulados com as interações que você realiza com eles.

Capacitado com o uso do DM, será possível identificar novos conhecimentos, transformando os dados contidos no banco de dados em informações importantes para

relacionamento com a administração. Com os resultados obtidos através desse processo, poder-se-á resolver problemas complexos e tomar decisões mais inteligentes e eficazes.

METODOLOGIA

As seções abordadas nesse capítulo procuram demonstrar todos os passos executados para a realização dessa pesquisa empírica.

Segundo Braga (2004), analistas de mineração de dados desenvolvem dois tipos de modelos: descritivos e preditivos. Na primeira etapa do trabalho, realiza-se a ADE dos dados dos alunos dos quatro Colégios Militares. Na segunda análise, utilizam-se AA e ACP aplicadas apenas nos dados do CMSM e do CMC, por estes utilizarem o módulo de controle de comportamento do SGE.

A terceira análise refere-se à predição de ocorrências. Baseado num modelo gerado a partir da ADI, utilizam-se dados do CMSM e do CMC, semelhante à análise anterior, mas acrescida de algumas variáveis.

1.12 Coleta de dados

Os dados foram coletados do SGE e estão armazenados em sistemas gerenciadores de bancos de dados como Oracle 7.4 e PostgreSQL 7.

As informações foram transpostas para tabelas simples com a utilização de programas específicos desenvolvidos apenas para esta necessidade. Esses programas objetivavam conectar o aplicativo a base de dados, podendo esta ser Oracle ou PostgreSQL, e importar os dados para um arquivo, no formato padrão, com a finalidade de ser aberto pelo programa de análises estatísticas.

Dessa forma, utilizam-se alguns programas como o Statistica, Microsoft Excel, SPSS Clementine e SAS.

1.13 Preparação dos dados

Nos dados de cadastro dos alunos do CMRJ, algumas informações não estão atualizadas para todos os alunos. Logo, foram utilizados somente os dados dos alunos com cadastro completo, totalizando 1276 alunos, cerca da metade dos alunos do efetivo total dessa instituição. Para os demais Colégios, foram utilizados todos os casos, pois o cadastro estava ajustado.

Salienta-se aqui que foi realizada uma filtragem nos registros utilizados, de maneira que foram excluídos os casos com valores nulos ou em branco, ou seja, alunos que tinham o campo grau em alguma disciplina em branco, por motivo de desligamento, estes foram apagados.

1.14 Análise descritiva

O rígido controle das informações em uma instituição de ensino militar, por si só, garante a confiabilidade e segurança dos dados utilizados nesta pesquisa, assegurando o que afirma Hayes (2001) que os dados obtidos refletem as informações válidas e viáveis.

Para a ADE, utilizam-se as variáveis:

- Média Geral da Série (MGS);
- Tipo do Amparo (TA)

A MGS é uma variável numérica, calculada a partir da média aritmética das disciplinas que o aluno realizou no ano de 2004, antes de realizar a recuperação final. Utiliza-se, no SCMB, a média de aprovação igual ou superior a cinco. Nesse caso, a quantidade de alunos com média inferior a cinco representa o número de alunos com baixo rendimento.

O TA é uma variável categórica, podendo ser Amparado, Concursado ou Transferido. Dar-se-á ênfase aos alunos Amparados e Concursados. Concursados são os alunos que ingressaram no SCMB através de concurso de admissão. Amparados são os alunos dependentes de militares que, por razões previstas nas leis de ensino do Exército Brasileiro, têm direito ao acesso a um Colégio Militar.

Quando se trabalha com variáveis qualitativas, precisa-se, primeiramente, transformá-las em variáveis quantitativas, para, só depois, aplicar a técnica desejada. Nessa análise tem-se a variável quantitativa Média Geral da Série (MGS) e transforma-se para uma menção, a qual é comparada com a origem do aluno.

Neste caso, realizou-se um estudo acerca da população analisada, a qual é formada pelos alunos do CMSM, CMRJ, CMC e CMBH, onde se procura identificar um padrão entre as características analisadas e traçar um modelo, definindo assim o perfil das escolas em estudo.

Primeiramente, procura-se apresentar o número de indivíduos de cada instituição e o número de elementos de acordo com a origem. Para melhor entendimento, apresenta-se o valor das menções utilizadas na transformação da variável MGS em menção de acordo com os padrões do Exército Brasileiro.

Nessa análise descritiva, procura-se identificar um comportamento por meio do cruzamento de variáveis e, a partir daí, caracterizar as instituições em estudo de acordo com os parâmetros gerados. Assim, apresenta-se, nesta primeira análise, a identificação do perfil dos Colégios com relação ao rendimento e a origem dos alunos.

Dessa forma, prossegue-se o estudo com a identificação da relação entre outras variáveis, como o comportamento dos alunos e seu rendimento nas disciplinas. Com a intenção de caracterizar as escolas em estudo, procede-se com uma análise individual dos Colégios para verificar qual apresenta maior evidência dessa relação.

1.15 Análise de aglomeração e componentes principais

Nesta etapa, aplicam-se os métodos multivariados. Dentre eles, destaca-se a análise fatorial, à qual deu-se maior ênfase por ser uma técnica estatística bastante eficaz, no que tange a trabalhos na área de pesquisa.

Para a AA e ACP utilizam-se variáveis numéricas referentes aos graus alcançados pelos alunos nas disciplinas da terceira série do ensino médio de 2004 e o comportamento dos alunos no final do ano letivo, dia 30 de novembro de 2004:

- Grau de Comportamento (GrauComp);
- Biologia (Bio);
- Educação Física (EF);
- Física (Fis);
- Geografia (Geo);
- História (Hist).
- Língua Estrangeira Moderna (LEM);

- Literatura (Lit).
- Metemática (Mat);
- Língua Portuguesa (Port);
- Química (Qui).

O Grau de Comportamento é um valor que se altera durante a permanência do aluno no Sistema Colégio Militar do Brasil. Quando o aluno ingressa num Colégio, recebe oito no grau de comportamento, mas, de acordo com as punições e/ou melhorias, esse valor é alterado.

Nesse caso, elabora-se uma tabela com as médias e desvio padrão das variáveis; depois, apresenta-se o inter-relacionamento das variáveis através da matriz de correlação.

Como existem correlações altas, procede-se, com a verificação dos grupos formados pelas variáveis, para identificar quais variáveis pertencem ao mesmo agrupamento e as variáveis mais importantes em cada fator.

Procede-se, então, com a ACP, onde se verifica os autovalores, as variâncias explicadas por cada autovalor, os autovalores acumulados e suas respectivas variâncias acumuladas, calculados a partir da matriz de correlação.

Com a extração das componentes principais do conjunto das variáveis em estudo, pode-se trabalhar com apenas dois caracteres para sua representação num plano, mas quando esse número aumenta, torna-se complicado analisar a nuvem de pontos formada. Para isso, utiliza-se a ACP, por se tratar de um método de redução do número de caracteres, e por permitir a representação geométrica dos indivíduos e dos caracteres, sendo que essa redução só é possível se os caracteres iniciais não forem independentes e possuírem coeficientes de correlação não-nulos.

Parte-se, então, para a identificação do número de componentes a serem definidos para a análise. A seleção de componentes é realizada de acordo com os valores próprios superiores à unidade.

A interpretação e análise dos resultados tornam-se o item de maior relevância no estudo, pois uma má interpretação acarreta perda de sentido da pesquisa. Portanto, após a seleção e a identificação das componentes a serem analisadas, é realizado um estudo de correlação entre as variáveis originais e a componente, possibilitando encontrar a variável que possuir maior influência naquela componente.

Na interpretação dos resultados, podem ser encontradas duas ou mais componentes com o mesmo grau de explicação, devendo-se obedecer a uma ordem de hierarquia. Para solucionar esse conflito, deve-se optar pelo maior autovalor que originou a componente extraída.

Estimam-se os autovetores e verifica-se a importância de cada variável através da matriz de correlações entre as componentes e as variáveis em estudo. Analisam-se os grupos formados pelo processo de agrupamento.

Esse estudo poderia seguir para uma análise individual dos quatro Colégios (CMC, CMSM, CMBH, e CMRJ), semelhante ao procedimento admitido na ADE, onde seria possível verificar qual instituição se adapta melhor ao padrão formado pelas AA e ACP. Optou-se por verificar a relação de alguns alunos com os fatores identificados.

Essa técnica reduz o número de variáveis originais, fornecendo um melhor entendimento do conjunto de dados. Além disso, possibilita ao pesquisador reduzir e sumarizar os dados, uma vez que examina todo o conjunto, sem a preocupação de verificar quais variáveis são dependentes ou independentes.

Logo, tal técnica possibilita somente verificar as relações de interdependência entre as variáveis analisadas, fornecendo subsídios para a administração avaliar o desempenho em relação ao comportamento, além de possibilitar um melhor entendimento sobre os casos.

Após a projeção das variáveis no plano fatorial e a identificação de alguns alunos para teste, gera-se a projeção dos casos no plano para localizar estes alunos, e identificam-se os eixos fatoriais e a as variáveis mais significativas.

Seguindo as análises dos planos fatoriais, é realizado um estudo final de todos os itens envolvidos no processo, e, a partir deste estudo, deve-se tirar as conclusões que são consideradas pertinentes, auxiliando, desta forma, na tomada de decisão e permitindo a concentração de esforços somente naqueles itens que necessitem de uma atenção especial.

1.16 Análise discriminante

Para a ADI utilizam-se as variáveis:
- Pontos Perdidos (PPerd);
- Grau de Comportamento (GrauComp);
- Situação da Matrícula (Situac);
- Biologia (Bio);
- Educação Física (EF);
- Física (Fis);
- Geografia (Geo);
- História (Hist).
- Língua Estrangeira Moderna (LEM);
- Literatura (Lit).
- Metemática (Mat);
- Língua Portuguesa (Port);
- Química (Qui).

A variável Pontos Perdidos (PPerd) representa o número de faltas obtidas durante o ano letivo de 2004. A variável categórica Situação da Matrícula (Situac) armazena três possíveis valores: Aprovado, Aprovado com PR e Reprovado.

Procura-se determinar quais disciplinas são mais importantes para a questão da aprovação final. Por se tratar de um método de classificação de casos, usa-se, nesta etapa do estudo, a ADI.

A seguir, são coletados dados individuais dos elementos de cada grupo. Neste caso, utiliza-se a variável categórica Situação (Situac) para se classificar os alunos e gerar a função discriminante, função de classificação e matriz de classificação.

Após, cria-se uma planilha no *Microsoft Excel*, com um sistema para entrada de dados nas variáveis mais significativas, e o mesmo apresenta as distâncias calculadas para cada situação de matrículas. Assim, pode-se identificar em qual classe se enquadra o suposto aluno testado no modelo de 2004.

Dessa forma, procede-se com três tipos de análise, envolvendo todas as variáveis de rendimentos e parte disciplinar dos alunos de quatro Colégios do SCMB.

RESULTADOS

Não se pode só considerar a escola como o único local onde a educação acontece, mas é o mais conhecido. A escola é mais visada e recebe investimentos para melhorias de profissionais, qualidade de ensino e aprendizagem.

A educação nas escolas busca proporcionar condições para que o cidadão absorva conhecimentos e construa uma vida digna, integrando-se em uma sociedade instruída e que tenha uma visão crítica das situações que o cercam.

Porém, hoje, o ensino ainda está fortemente relacionado com a avaliação e a "nota", que surge com este processo. Novas idéias e propostas estão surgindo, outras até estão em prática, mas é um processo difícil de ser mudado. Hoje em dia, instituições dizem fazer uma avaliação construtivista, conceituando alunos de maneira diferenciada, mas no fim a conceituação sempre é representada através de um grau.

Existe também a preocupação, principalmente governamental, em reduzir o índice de reprovações nas escolas públicas, mas, algumas vezes, é deixada de lado a qualidade do ensino oferecido. Sempre que se fala em educação, os principais itens analisados são os índices de aprovação e reprovação. Há quem pense que a melhor escola é aquela que mais aprova.

Mas esse assunto promove novas discussões, que poderão servir para estudo em outra ocasião. O interesse da análise dos dados apresentados a seguir é objetivado em relação às escolas de nível fundamental e médio, relacionando aprovações e reprovações nessas instituições analisadas.

Com o objetivo de conhecer o comportamento das variáveis, desenvolve-se um estudo de caráter descritivo, seguido da aplicação de análises multivariadas.

1.17 Análise descritiva

Um projeto de exploração em banco de dados destina-se a duas finalidades: traçar o perfil das pessoas/instituições e predizer algum resultado baseado nos dados armazenados.

Trata-se por perfil do rendimento de ensino, a identificação de características individuais ou semelhantes, baseado em interpretações capturadas nos dados adquiridas ao longo do tempo. Para traçar o perfil dos alunos e dos Colégios em estudo, inicialmente aplica-se uma análise descritiva, na qual a população em estudo é composta por 3360 alunos dos quatro Colégios Militares. A Tabela 01 apresenta os número de indivíduos de cada instituição.

TABELA 01 – Número de elementos por colégio.

Colégio	Número de Alunos
CMBH	695
CMC	725
CMRJ	1276
CMSM	664
Total	3360

Nesta análise, procura-se relacionar o rendimento com a origem do aluno, onde o rendimento é representado pela variável Média Geral da Série (MGS). Na Tabela 02 verifica-

se uma equivalência entre os alunos amparados e concursados, objetos principais desta análise.

TABELA 02 – Número de elementos por origem.

Origem	Número de Elementos
Amparado	1798
Concursado	1507
Transferido	55
Total	3360

O desvio padrão e o erro padrão da média são medidas calculadas em torno da média e a intenção é quantificar a variabilidade dos dados em torno da média. Por ser, então, uma medida relativa, deve existir uma referência para que se faça a interpretação. A média da variável MGS foi de 7,21 e o desvio padrão de 1,20. Esses valores mostram uma concentração em torno da média, pois o coeficiente de variação de Pearson foi de 16,6 %. Isso indica que a média é representativa do conjunto de dados em estudo.

A Tabela 03 apresenta a categorização do valor da média geral de série, para a sumarização da análise. Classificam-se as médias de acordo com as menções estabelecidas pelo ensino militar, para melhor interpretação dos resultados. A menção insuficiente (I) representa uma média abaixo de cinco, caracterizando um mau rendimento.

TABELA 03 – Categorização do valor da média.

Descrição	Menção	Menor Valor	Maior Valor
Excelente	E	9.5	10
Muito Bom	MB	8.0	9.49
Bom	B	6.0	7.99
Regular	R	5.0	5.99
Insuficiente	I	0	4.99

Procede-se com a criação de um modelo para comparação das características individuais de cada instituição. O gráfico da Figura 07, que representa todos os Colégios Militares, apresenta uma concentração maior de alunos concursados com rendimento bom e muito bom, enquanto os alunos amparados concentram-se no rendimento bom. Ainda se nota que o rendimento abaixo da média cinco, ou seja, com menção insuficiente, encontra-se apenas nos alunos amparados.

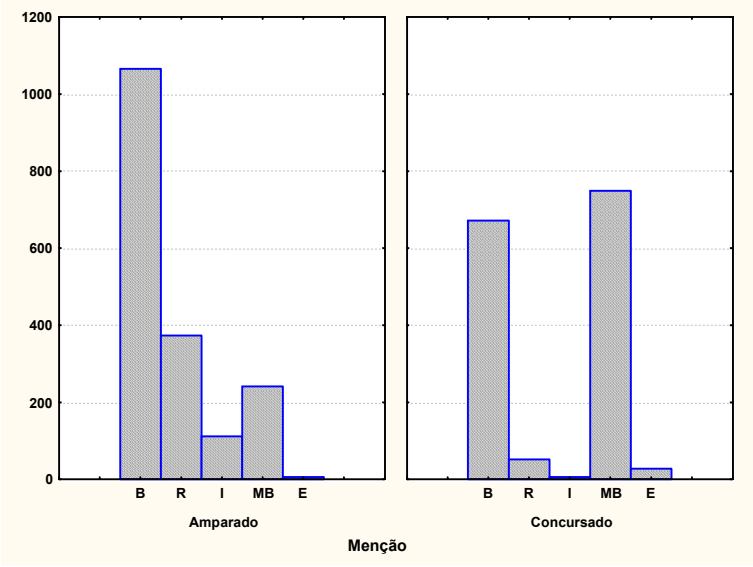


FIGURA 07 – Gráfico de Colunas das origens traçadas em relação ao rendimento de todos os Colégios Militares.

Existe aproximadamente o mesmo número de alunos amparados e concursados. Isso mostra que os alunos concursados apresentam melhor desempenho considerando a média global da série. Com a intenção de caracterizar as escolas em estudo, procede-se com uma análise individual dos Colégios, para verificar qual apresenta maior semelhança com a característica evidenciada.

A Figura 08 refere-se ao CMC e apresenta um comportamento semelhante aos gráficos da Figura 07.

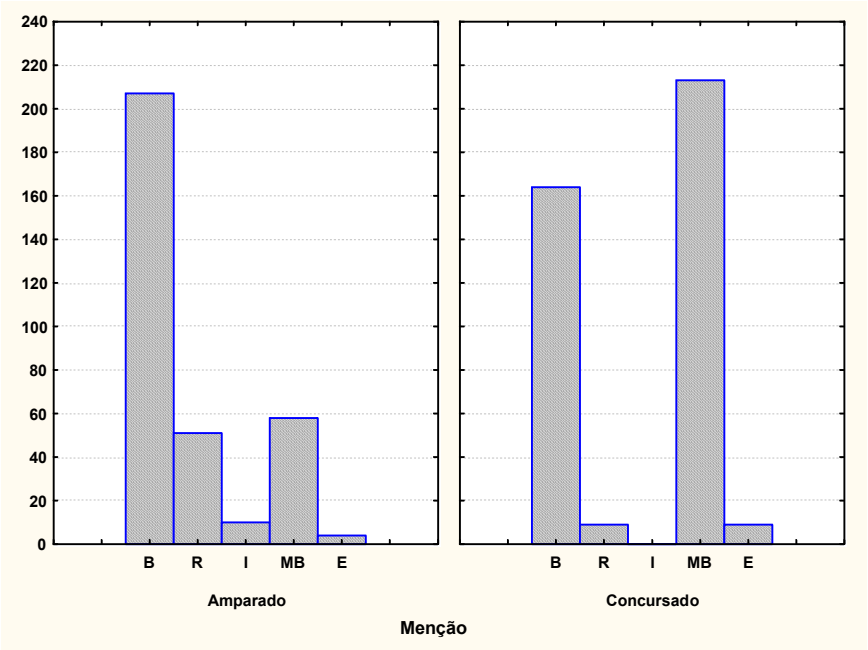


FIGURA 08 – Gráfico de Colunas das origens do CMC traçados em relação ao rendimento.

No gráfico dos amparados, nota-se uma baixa proporção de alunos com menção Insuficiente (I) em relação às menções Muito Bom (MB) e Bom (B). Da mesma forma, os outros Colégios (CMSM e CMBH) também apresentam comportamento semelhante ao do CMC. Contudo; no gráfico da Figura 09, que representa o CMRJ, nota-se uma maior

proporção de alunos com menção insuficiente na classe dos amparados. Isso comprova um maior número de alunos com rendimento baixo nos amparados, principalmente no CMRJ.

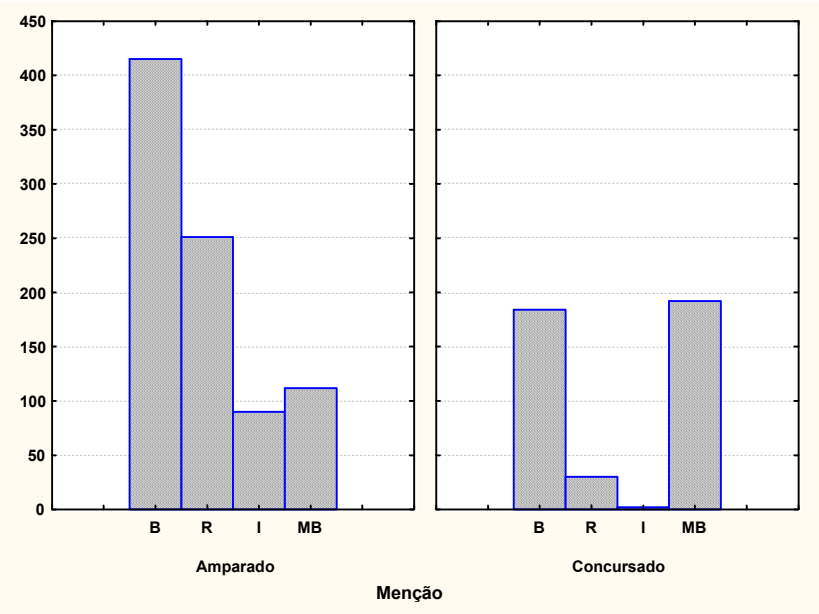


FIGURA 09 – Gráfico de Colunas das origens do CMRJ traçados em relação ao rendimento.

Nesta ADE, procura-se identificar um padrão para, a partir daí, caracterizar as instituições em estudo, de acordo com o modelo criado no início. Assim, apresenta-se, nesta primeira análise, a identificação do perfil dos Colégios com relação ao rendimento e a origem dos alunos. Esta análise torna-se proveitosa no momento em que se pode identificar as características individuais de cada instituição.

Dessa forma, prossegue-se o estudo com a identificação da relação entre outras variáveis, como o comportamento dos alunos e seu rendimento nas disciplinas.

1.18 Caracterização do CMSM e CMC em relação aos rendimentos de ensino e comportamento

Para esta análise, utilizam-se os dados de comportamento do CMSM e CMC, armazenados no Sistema de Gestão Escolar (SGE), porque apenas estes utilizam o módulo de controle de comportamento.

Aqui, procura-se identificar a relação entre o grau de comportamento e o rendimento escolar, considerando-se as disciplinas da 3ª série do Ensino Médio.

Os valores das médias e desvios padrão das variáveis, utilizando todos os casos deste estudo, estão representados na Tabela 04.

TABELA 04 – Médias e desvio padrão das variáveis.

Variáveis	Número de Alunos	Média	Desvio Padrão	Mínimo	Máximo
GrauComp	184	9,433	1,165	3,400	10,000
Bio	184	7,137	0,992	5,000	9,200
EF	184	7,671	1,676	0,900	10,000
Fis	184	7,229	0,978	5,300	9,800
Geo	184	7,398	0,751	5,700	9,400
Hist	184	7,902	1,019	5,100	9,700

LEM	184	7,293	0,997	4,900	9,500
Lit	184	6,848	0,917	4,700	9,700
Port	184	7,125	0,831	5,100	9,700
Mat	184	6,841	1,304	3,400	9,700
Qui	184	7,132	0,999	5,200	9,500

Observa-se que a maior média é do grau de comportamento (9,4), seguido de História (7,9) e Educação Física (7,7). As médias concentram-se na faixa de 6,8 a 7,4.

De acordo com a matriz de correlação, apresentada na Tabela 05, que mostra o inter-relacionamento das variáveis, verifica-se uma baixa correlação das disciplinas com o grau de comportamento.

TABELA 05 – Matriz de correlação entre as variáveis.

Variáveis	GrauComp	Bio	EF	Fis	Geo	Hist	LEM	Lit	Port	Mat	Qui
GrauComp	1,000										
Bio	0,146	1,000									
EF	0,221	0,137	1,000								
Fis	0,231	0,671	0,255	1,000							
Geo	0,170	0,608	0,236	0,667	1,000						
Hist	0,158	0,420	0,295	0,569	0,623	1,000					
LEM	0,128	0,541	0,093	0,566	0,540	0,528	1,000				
Lit	0,214	0,621	0,163	0,653	0,694	0,702	0,615	1,000			
Port	0,217	0,759	0,108	0,686	0,594	0,415	0,570	0,685	1,000		
Mat	0,273	0,692	0,179	0,742	0,558	0,415	0,515	0,542	0,706	1,000	
Qui	0,249	0,682	0,211	0,788	0,641	0,544	0,542	0,611	0,658	0,773	1,000

A AM é uma ferramenta estatística capaz de revelar o comportamento conjunto de variáveis quando se faz uma análise simultânea de diversas características de um objeto ou caso. Também ela é útil quando existe um grau de associação significativo entre as variáveis.

A única disciplina que não apresentou forte correlação com as demais foi Educação Física (EF). Nas demais disciplinas, existe uma forte correlação entre as variáveis, o que comprova a afirmação de que um aluno que apresenta um bom desempenho em uma disciplina também apresenta nas outras, mas não significa que ele tenha um bom comportamento ou bom rendimento em Educação Física.

Como existem fortes correlações, procede-se, então, com a verificação dos grupos formados pelas variáveis, para identificar quais variáveis pertencem ao mesmo agrupamento.

Parte-se para a AC, definida no item 2.3.1, que tem por objetivo analisar a proximidade geométrica entre as variáveis estudadas, sendo utilizada sempre que se quer identificar grupos de características semelhantes, levando-se em conta todas as medidas originais. Nesse caso, utiliza-se o processo de aglomeração hierárquica, com o método de variância de Ward.

A Figura 10 mostra o comportamento do dendograma com todas as variáveis, na qual pode-se identificar a formação de dois grupos, os quais possuem as variáveis de maior relevância dentro do conjunto.

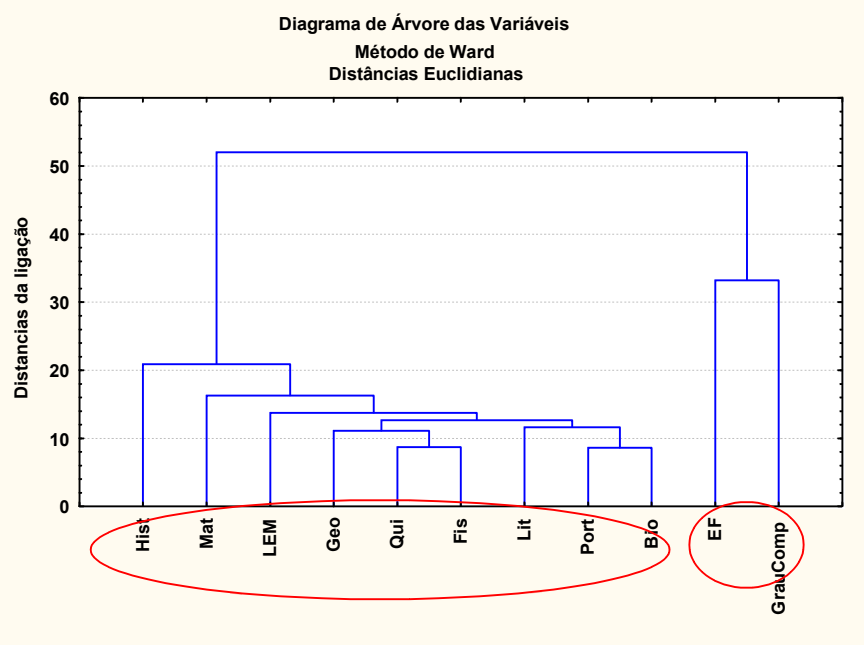


FIGURA 10 – Dendograma envolvendo as variáveis em estudo.

O primeiro agrupamento é formado pelas variáveis Grau de Comportamento (GrauComp) e Educação Física (EF), o segundo, pelas demais disciplinas. Identifica-se um agrupamento que representa os atributos da área psicomotora/afetiva e outro formado pelas áreas de ciências, que exigem estudo, escrita e leitura.

Nota-se ainda que o grupo formado por Língua Portuguesa (Port) e Biologia (Bio), assim como o grupo formado por Química (Qui) e Física (Fis) estão juntos porque apresentam médias semelhantes, ou seja, um aluno que tem bom rendimento em uma disciplina também apresenta esta característica na outra disciplina do grupo.

Após a identificação das variáveis por meio de aglomerações, procede-se com a ACP para identificar as variáveis mais importantes em cada fator. A Tabela 06 apresenta os autovalores, as variâncias explicadas por cada autovalor, os autovalores acumulados e suas respectivas variâncias acumuladas, calculados a partir da matriz de correlação mostrada na Tabela 04.

O percentual de variância explicada pelos dois primeiros autovalores é de 65,617%, que representa a variabilidade total do sistema. Parte-se, então, para a identificação do número de fatores a serem definidos para a análise.

TABELA 06 – Autovalores e percentual de variância explicada.

Fatores	Autovalores	Variância Explicada (%)	Autovalores Acumulados	Variância Explicada Acumulada (%)
1	6,080	55,269	6,080	55,269
2	1,138	10,348	7,218	65,617
3	0,938	8,530	8,156	74,147
4	0,732	6,651	8,888	80,798
5	0,479	4,358	9,367	85,156
6	0,447	4,066	9,814	89,222
7	0,331	3,007	10,145	92,229

8	0,258	2,350	10,404	94,579
9	0,224	2,034	10,627	96,613
10	0,195	1,775	10,823	98,387
11	0,177	1,613	11,000	100,000

Observa-se que esta explicação é devida aos autovalores superiores a um, onde o 1º autovalor é 6,080 e o 2º é 1,138, o qual pode ser corroborado através do método gráfico, sugerido por Cattel (1966), visualizado na Figura 11.

Este gráfico, utilizado para a ACP, representa a porcentagem de variação explicada pela componente nas ordenadas, e os autovalores em ordem decrescente nas abscissas. Este critério, considera as componentes anteriores ao ponto de inflexão da curva. Nesse caso, observa-se uma quebra busca após a 2ª componente principal, sugerindo-se que sejam consideradas apenas as duas primeiras componentes principais.

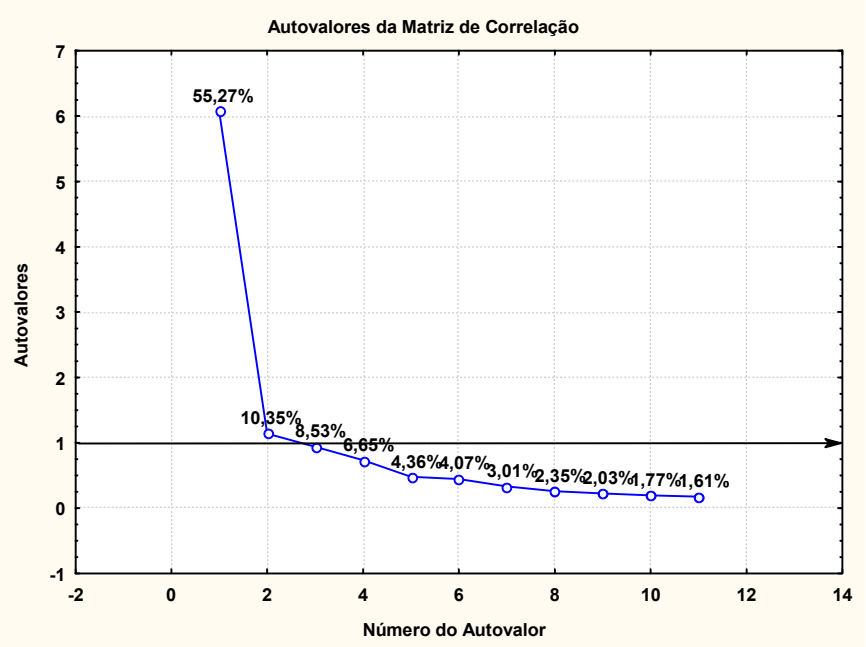


FIGURA 11 – Gráfico de declive dos autovalores.

Outro método, para a seleção, é a proporção da variação explicada pelas componentes, onde os três primeiros componentes explicam 74,147% da variabilidade total, portanto, mais que 70%, que é o critério adotado para a escolha das componentes a serem utilizadas. O critério da escolha do autovalor maior que um e o gráfico da Figura 03 corroboram para a indicação do método em que devem ser usadas apenas as duas primeiras componentes para uma avaliação das variáveis. Utilizaram-se os autovalores da Tabela 06, estimaram-se os autovetores para escrever a combinação linear que dará origem aos fatores, conforme evidenciado na Tabela 07.

TABELA 07 – Autovetores.

Variável	Fator 1	Fator 2
GrauComp	-0,122	-0,575
Bio	-0,331	0,193
EF	-0,114	-0,735
Fis	-0,355	-0,010
Geo	-0,328	-0,013
Hist	-0,287	-0,162
LEM	-0,294	0,168

Lit	-0,338	0,043
Port	-0,337	0,180
Mat	-0,333	0,048
Qui	-0,349	0,009

Observa-se, ainda, que a maioria das variáveis são significativas para o Fator 1 e seus valores giram em torno de 0,30 a 0,35. Logo, opta-se por analisar somente dois fatores, transformando um problema de dimensão 11 para o plano. Para a composição de cada fator, verifica-se a importância de cada variável através da matriz de correlações apresentadas na Tabela 08.

TABELA 08 – Correlação entre os fatores e as variáveis.

Variável	Fator 1	Fator 2
GrauComp	-0,300	-0,613
Bio	-0,817	0,206
EF	-0,280	-0,785
Fis	-0,875	-0,011
Geo	-0,809	-0,014
Hist	-0,708	-0,173
LEM	-0,726	0,179
Lit	-0,833	0,045
Port	-0,830	0,192
Mat	-0,820	0,051
Qui	-0,860	0,010

Os resultados mostram a relação das componentes principais com as variáveis originais, as quais, quando significativas a um nível de 7%, são destacadas. Isso não impede a realização de um estudo da correlação entre as componentes principais e os dados originais, para salientar quais são as variáveis mais representativas.

A AF, por ser uma técnica aplicada para identificar fatores num determinado conjunto de medidas realizadas, é utilizada, também, como uma ferramenta, em tentativas para reduzir um grande conjunto de variáveis para um conjunto mais significativo, representado pelos fatores. Esse método determina quais variáveis pertencem a quais fatores e o quanto cada variável explica cada fator.

Deve-se buscar uma interpretação física para melhor entender esses construtos. Depois de definidos os fatores de estudo, representam-se graficamente, na Figura 12, as variáveis no plano fatorial para comprovar os agrupamentos formados.

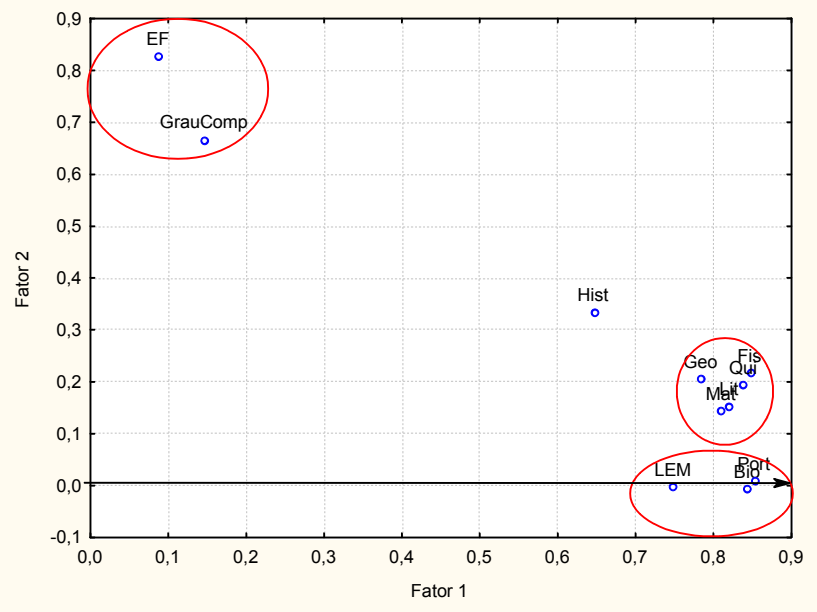


FIGURA 12 – Plano Fatorial – Fator 1 x Fator 2.

Nota-se que os agrupamentos são semelhantes aos formados na AC, representando o fator um como áreas das ciências, que exigem estudo, escrita e leitura, e o fator dois os atributos da área psicomotora/afetiva.

Ainda, Língua Portuguesa (Port) e Biologia (Bio) aparecem no mesmo grupo, assim como Química (Qui) e Física (Fis). Grupos, estes, formados na AC, corroborando para a semelhança entre as variáveis.

Nota-se, aqui, um maior distanciamento de História das demais disciplinas formadoras do agrupamento da área das ciências, o que expressa menor semelhança das médias de História com as das demais disciplinas.

Este estudo poderia seguir para uma análise individual dos Colégios, semelhante ao procedimento admitido na ADE, onde seria possível verificar qual instituição se adapta melhor ao padrão formado pelas AC e ACP.

Optou-se por verificar a relação de alguns alunos com os fatores identificados.

A Figura 13 apresenta a projeção das variáveis no círculo unitário.

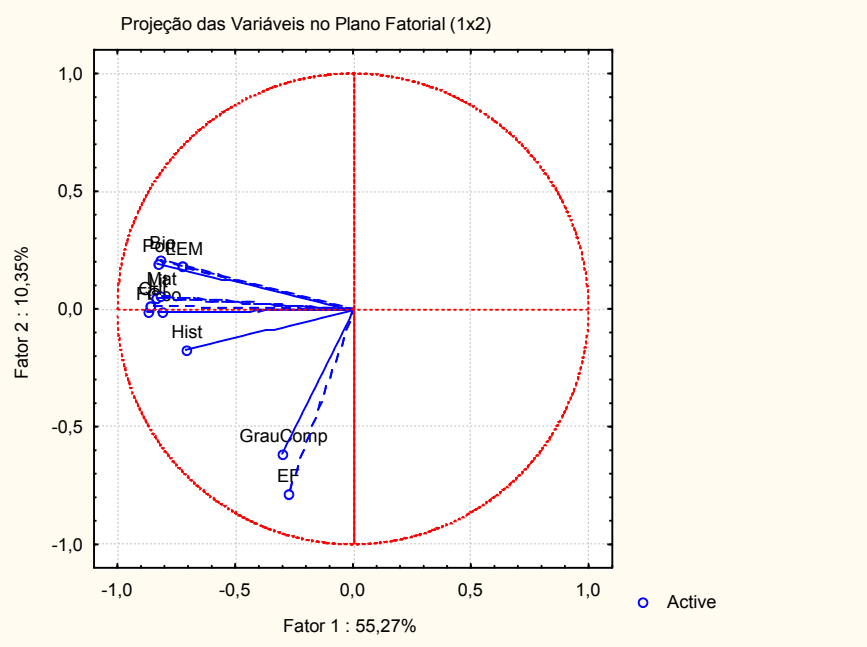


FIGURA 13 – Projeção das variáveis no círculo unitário.

Observa-se, visualmente, neste gráfico, o que foi identificado na Figura 12. Devido à comprovação destes grupos formados, Foram escolhidos seis alunos, três de cada Colégio. Os valores das variáveis, para os casos escolhidos, estão apresentados na Tabela 09.

TABELA 09 – Alunos escolhidos para visualização no círculo unitário.

	Colégio	GrauComp	Bio	EF	Fis	Geo	Hist	LEM	Lit	Port	Mat	Qui
Aluno1	CMSM	10	9,2	9,8	9,2	8,7	9,2	9,1	8,3	8,6	9,7	8,7
Aluno2	CMSM	10	9,2	8,5	8,8	8,5	9,1	8,8	8,2	8,1	9,2	8,5
Aluno3	CMC	10	9,2	8,5	9,0	9,0	9,6	8,3	9,1	9,3	8,9	8,6
Aluno4	CMC	10	9,2	8,5	9,8	9,2	9,7	8,9	9,7	9,7	9,2	9,3
Aluno5	CMSM	6,4	8,0	0,9	6,5	7,1	7,8	9,4	7,5	7,5	5,7	6,8
Aluno6	CMC	10	5,9	10	6,5	6,4	7,6	6,5	5,4	6,1	5,9	7,5

Na Figura 14, visualizam-se os alunos no plano fatorial.

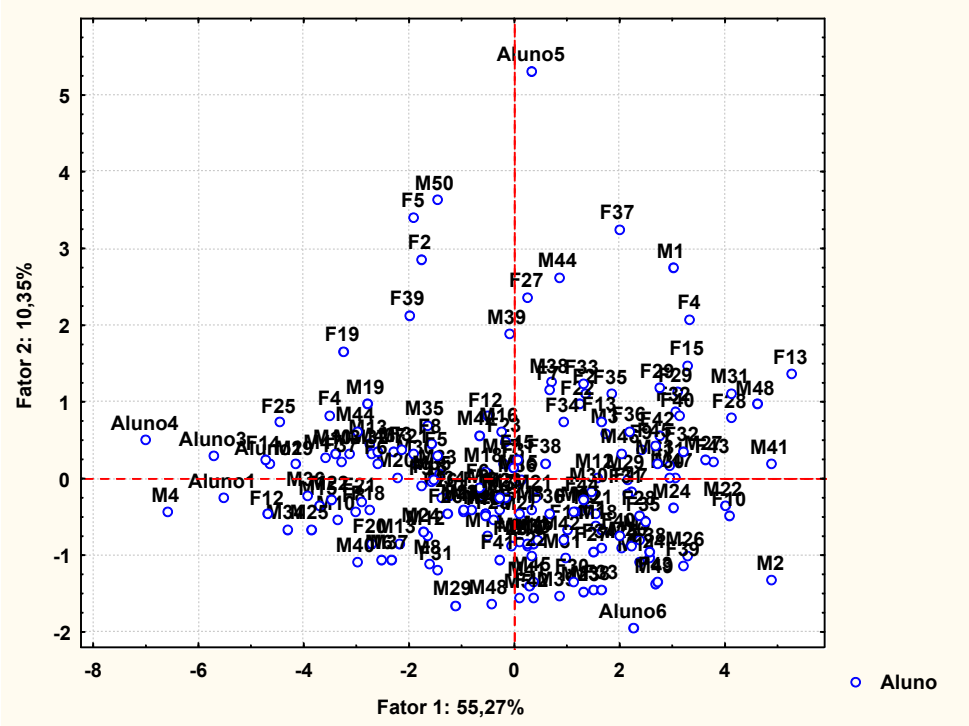


FIGURA 14 – Projeção dos casos no plano fatorial.

Observa-se que, dos alunos selecionados para análise, o Aluno1, Aluno2, Aluno3 e Aluno4 estão no mesmo sentido das disciplinas da área das ciências na Figura 13.

Já o Aluno5 apresentou a menor nota de Educação Física e Comportamento Bom, abaixo da média geral de comportamento, o que determinou sua localização oposta à localização das disciplinas de Educação Física e Grau de Comportamento (GrauComp).

Dessa forma, utilizando-se AC, ACP e AF, pôde-se identificar um padrão entre os Colégios e classificar alunos de acordo com o modelo formado. Torna-se válida a análise, pois se pode caracterizar o perfil desses alunos.

1.19 Análise discriminante

Após a identificação de que o Grau de Comportamento (GrauComp) não é significante em relação as demais variáveis, procura-se determinar quais disciplinas são mais importantes para a questão da aprovação final, ainda se utilizando outra variável que é a dos Pontos Perdidos (PPerd). Por se tratar de um método de classificação de casos, usa-se, nesta etapa do estudo, a ADI.

A seguir, são coletados dados individuais dos elementos de cada grupo. A ADI procura estimar a combinação linear das características individuais de cada elemento que melhor discrimina entre os grupos pré-estabelecidos.

Nesse caso, utiliza-se a variável categórica Situação (Situac) para se classificar os alunos e gerar a função discriminante. A Tabela 10 apresenta o resumo da análise da função discriminante, a qual determinou três variáveis significativas, por meio da regressão passo a passo (*stepwise*) com o método Eliminação para Trás (*Backward*), definido no item 2.3.4, onde o λ de Wilks foi superior a 0,65680 e p<0,0000.

Na Eliminação para Trás, inicialmente, todas as variáveis prognosticadoras são incluídas na equação de regressão. Removem-se, então, as prognosticadoras, uma de cada vez, com base na razão F.

TABELA 10 - Sumário do resultado da função discriminante.

Variável	λ de Wilks	p
Fis	0,680448	0,000038
Geo	0,688187	0,000001
Mat	0,685273	0,000005

As disciplinas selecionadas são as mais representativas no que se refere à classificação pela situação da matrícula. Isso significa que, no boletim do aluno, essas disciplinas são as que mais influenciaram na caracterização da situação de aprovação do aluno no ano de 2004, sem a realização da Prova de Recuperação (PR).

A Tabela 11 apresenta as funções de classificação para cada tipo de situação do aluno.

TABELA 11 – Funções de classificação.

Variável	Aprovado	Aprovado c/PR	Reprovado
Fis	1,14	0,02	0,76
Geo	7,94	6,65	7,72
Mat	0,73	0,73	-0,57
Constante	-36,86	-23,93	-28,63

Dessa forma, pode-se identificar a seguinte função de classificação para :

$Y_{APROVADOS} = 1,14 * Fis + 7,94 * Geo + 0,73 * Mat -36,86$

$Y_{APROVADOS\ C/PR} = 0,02 * Fis + 6,65 * Geo + 0,73 * Mat -23,93$

$Y_{REPROVADOS} = 0,76 * Fis + 7,72 * Geo -0,57 * Mat -28,63$

A Matriz de Classificação, apresentada na Tabela 12, demonstra o percentual de validação da função discriminante, onde se pode notar que, para os Aprovados, a função discriminante acerta em 98,4 % dos casos. Nota-se ainda que o percentual total de acerto do modelo é de 90,7 %.

TABELA 12 – Matriz de classificação.

	Percentual	Aprovado	Aprovado c/PR	Reprovado
Aprovado	98,42	499	8	0
Aprovado c/PR	52,83	25	28	0
Reprovado	4,54	17	4	1
Total	90,72	541	40	1

Observam-se alguns erros na classificação dos alunos, onde não se pode estimar sobre os Aprovados e Reprovados. Se o conjunto de dados fosse maior, esses valores poderiam apresentar maior confiabilidade.

Após a identificação das variáveis significantes, parte-se para uma aplicação prática, onde, informa-se o provável grau para as disciplinas selecionadas pela função discriminante, e apresenta-se um resultado gerado pela classificação. A Tabela 12 apresenta as médias das variáveis, para cada tipo de situação de matrícula.

TABELA 13 – Média das variáveis e situações de matrícula.

Variável	Situação	Média

Fis	Aprovado	6,92
Fis	Aprovado c/ PR	4,86
Fis	Reprovado	5,15
Geo	Aprovado	7,61
Geo	Aprovado c/ PR	5,87
Geo	Reprovado	6,39
Mat	Aprovado	7,03
Mat	Aprovado c/ PR	5,22
Mat	Reprovado	4,58

Utiliza-se, como exemplo um suposto aluno a ser testado no modelo criado. Informa-se para Matemática o grau igual 5,5, para Geografia, o grau igual a 6 e Física, o grau igual a 6. Para a classificação do aluno foi utilizada a distância de Mahalanobis, descrita no item 2.3.1.

A seguir, apresenta-se o procedimento para cálculo da situação dos Aprovados. O mesmo procedimento deve ser executado para os casos de Aprovação c/ PR e Reprovação.

Sendo (-0,91854 -1,60572 -1,53156) a matriz referente à diferença entre os valores informados para as variáveis e suas médias e, a matriz de covariância representada por:

$$S = \begin{pmatrix} 1,227427 & 0,675969 & 1,059297 \\ 0,675969 & 0,865983 & 0,707058 \\ 1,059297 & 0,707058 & 1,66003 \end{pmatrix}$$

Sua inversa é representada por:

$$S^{-1} = \begin{pmatrix} 2,164434 & -0,86137 & -1,01428 \\ -0,86137 & 2,113258 & -0,35044 \\ -1,01428 & -0,35044 & 1,398896 \end{pmatrix}$$

Utilizado-se a distância de Mahalanobis, tem-se:

$$D^2 = (-0,91854 \quad -1,60572 \quad -1,53156) * \begin{pmatrix} 2,164434 & -0,86137 & -1,01428 \\ -0,86137 & 2,113258 & -0,35044 \\ -1,01428 & -0,35044 & 1,398896 \end{pmatrix} * \begin{pmatrix} -0,91854 \\ -1,60572 \\ -1,53156 \end{pmatrix} = 3,44$$

Assim, obtém-se o resultado de 3,44 para a distância dos aprovados. Executando-se o mesmo procedimento para o caso dos Aprovados com PR, o valor ficou igual a 5,17. Para o caso dos Reprovados, o valor foi de 4,97.

Dessa forma, pode-se afirmar, com 98,42% de certeza, que o referido aluno foi classificado na situação Aprovado sem realizar recuperação no final do ano letivo, pois o menor valor da distancia é a dos Aprovados.

A Figura 15 apresenta os dados e os procedimentos utilizando o Excel.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Fis	Geo	Mat	Situac											
2	9	9	8,9	Aprovado	Med Fis Aprov		6,91854		Med Geo Aprov		7,60572		Med Mat Aprov		7,031558
3	5,4	7,4	5,3	Aprovado	Med Fis Aprov c P		4,85849		Med Geo Aprov c PR		5,873585		Med Mat Aprov c PR		5,224528
4	8	7,5	6,5	Aprovado	Med Fis Reprov		5,15		Med Geo Reprov		6,395455		Med Mat Reprov		4,581818
5	7,3	7,8	7	Aprovado											
6	8,7	8	8,3	Aprovado	Fis->	5,5	Geo ->	5	Mat ->	5	Situac ->				
7	8,5	7,5	6	Aprovado											
8	7,3	7,6	6,7	Aprovado											
9	7,7	7	6,8	Aprovado					Mat Cov				Mat Inv		
10	9,8	9,2	9,2	Aprovado	APROVADO				1,227427	0,6759689	1,059297		2,164434	-0,86137	-1,01428
11	8	8,5	7,5	Aprovado					0,675969	0,8659831	0,707058		-0,86137	2,113258	-0,35044
12	5,9	7,6	5,6	Aprovado	Difs				1,059297	0,7070581	1,66003		-1,01428	-0,35044	1,398896
13	6,7	7,9	5,7	Aprovado	-1,4185	-2,6057	-2,0316								
14	8,1	8,3	8,6	Aprovado	-2,6057				1,234731	-3,5727186	-0,489982		8,553415		
15	9,6	9,4	9,4	Aprovado	-2,0316										
16	6,7	7,5	6,2	Aprovado											
17	5,3	6,8	5,2	Aprovado	APROVADO COM PR				0,375635	0,2038092	-0,102189		3,928565	-2,00418	0,657976
18	7,5	8,9	7,5	Aprovado					0,203809	0,3857173	-0,041994		-2,00418	3,639831	-0,10774
19	8,4	8,2	6,7	Aprovado	Difs				-0,102189	-0,0419936	0,482229		0,657976	-0,10774	2,203755
20	9,3	8,1	8,6	Aprovado	0,64151	-0,8736	-0,2245								
21	7,7	8,1	7	Aprovado	-0,8736				4,123299	-4,4412102	0,021415		6,520101		
22	6,7	7,5	6,2	Aprovado	-0,2245										
23	8	7	6,9	Aprovado											
24	7,4	8	6,5	Aprovado	REPROVADO				0,638864	0,3243182	0,120909		2,848714	-1,92851	-1,60855
25	8,1	7,5	5,8	Aprovado					0,324318	0,4886157	-0,011446		-1,92851	3,354568	1,191883
26	9,2	7,6	8,2	Aprovado	Difs				0,120909	-0,0114463	0,227851		-1,60855	1,191883	5,302279
27	5,3	6,6	5,6	Aprovado	0,35	-1,3955	0,41818								
28	6,5	6,6	5,2	Aprovado	-1,3955				3,015537	-4,8577024	-0,008894		7,830421		
29	9,3	7,8	8,6	Aprovado	0,41818										
30	7,6	8,6	6	Aprovado											

FIGURA 15 – Planilha do Excel utilizada na análise discriminante.

Essa análise tornou-se útil para identificar as disciplinas mais significativas em relação à situação da matrícula e em qual classe se enquadra o suposto aluno testado no modelo de 2004.

CONCLUSÕES E SUGESTÕES

Utilizando técnicas estatísticas multivariadas, baseado no rendimento dos alunos, elaboraram-se alguns modelos de perfil dos Colégios e dos alunos. Nas três análises realizadas, verifica-se a relação entre alguns indicadores de qualidade, disponibilizando, assim, subsídios para a tomada de decisões da administração.

Na primeira análise, pode-se identificar um padrão entre os Colégios e classificar as escolas de acordo com o modelo formado, onde se conclui que os alunos concursados apresentam melhor desempenho que os amparados, considerando-se a média global da série. Constata-se, ainda, que há um maior número de alunos com rendimento baixo nos amparados, principalmente no CMRJ. A representação do rendimento, comparada com a origem do aluno, através de histogramas na ADE, disponibiliza uma visão clara das distribuições formadas, o que comprova o eficiente uso da técnica empregada.

Na segunda análise, verifica-se a relação entre as disciplinas e o comportamento, onde se caracterizam dois Colégios, e classificam-se os alunos de acordo com o modelo formado. Através da AA, pode-se identificar um agrupamento, que representa os atributos da área psicomotora/afetiva, e outro, formado pelas áreas de ciências/cognitivas.

Nota-se, ainda, um agrupamento das disciplinas de Língua Portuguesa e Biologia, assim como Química e Física. Esses estão agrupados porque apresentam médias semelhantes, ou seja, um aluno que tem bom rendimento em uma disciplina, também apresenta esta característica na outra disciplina do grupo.

Usa-se ACP, por ser uma técnica utilizada na tentativa de reduzir um grande conjunto de variáveis para um conjunto mais significativo, representado pelos fatores, onde se nota que os agrupamentos formados são semelhantes aos formados na AA. Utiliza-se a ACP para identificar as variáveis mais importantes em cada fator.

Com a intenção de verificar a relação de alguns alunos com os fatores identificados, classificam-se seis alunos de acordo com o modelo formado. Torna-se válida a análise, pois se pode caracterizar o perfil desses alunos em relação aos graus obtidos nas disciplinas e o comportamento.

Na terceira análise, através da ADI, identifica-se que as disciplinas de Física, Matemática e Geografia são as mais representativas no que se refere à classificação pela situação da matrícula e, ainda, que essas disciplinas são as que mais influenciaram na caracterização da situação de aprovação do aluno, no ano de 2004. Desta forma, cria-se um modelo para caracterizar um tipo de perfil para aprovação, e utiliza-se, como exemplo, um suposto aluno com seus graus nas disciplinas mais significativas.

Assim, pode-se afirmar que o referido aluno foi classificado na situação Aprovado sem realizar recuperação no final do ano letivo. Não é o ideal para predição de acontecimentos, mas pode-se admitir que um aluno que se enquadra no perfil de aprovação em 2004 provavelmente terá um bom rendimento em 2005, seguindo uma uniformidade dos modelos gerados a cada ano.

Nesse caso, a técnica foi válida porque se pode classificar alunos em situações de aprovação, relacionando-os com o rendimento de ensino. Sugere-se a aplicação de análise de regressão para poder predizer situações de aprovação, ou reprovação, de alunos.

Esta pesquisa é importante para os Colégios Militares pois, utilizando-se informações sumarizadas e correlacionadas, representadas graficamente, o comando das instituições adquire maior dinamismo no controle dos processos de ensino. Através do detalhamento das técnicas estatísticas aplicadas na exploração de dados, pode-se conhecer melhor a análise

multivariada, no sentido de fornecer informações baseadas em ferramentas tecnológicas, para a tomada de decisões.

A utilização de indicadores de qualidade, armazenados em bancos de dados, defendida por Gil (1992), representa uma necessidade para os órgãos públicos. Visando à descoberta de conhecimento nessas bases, deve-se ampliar a estrutura de dados das instituições para armazenar indicadores sócio-econômicos, atributos da área afetiva, dados médicos e psicológicos e índices de satisfação das pessoas.

As novas ferramentas de DM possuem ambientes gráficos, onde se modela um projeto de exploração de dados. Este projeto, conectado com um banco de dados dinâmico, mostra cenários pré-definidos em tempo real, podendo ser acompanhado ao longo do tempo.

Logo, sugestiona-se a utilização de uma ferramenta de controle estatístico nas instituições, para determinação das características dinâmicas dos processos que envolvem a área de ensino. Amparado na significância das informações contidas nas imensas bases de dados, estes, incluídos no decorrer da existência da escola, os projetos de exploração devem ser definidos pela administração de ensino, determinando que indicadores analisar.

Cinco tipos de conhecimento são fundamentais para um bom trabalho de exploração de dados: conhecimento dos dados analisados, conhecimento na área da qualidade, conhecimento em estatística, conhecimento dos programas de computador com recursos de mineração de dados, e, principalmente, conhecimento das regras do negócio.

É imprescindível dispor de analistas capacitados que saibam interagir com os sistemas, de forma a conduzí-los para uma extração de padrões úteis e relevantes.

Objetivando aumentar competência e a criatividade nas instituições no que se refere à organização e gestão de sistemas de qualidade, através da metodologia desenvolvida neste trabalho, pode-se aplicar essas análises em instituições de ensino público e/ou privado, caracterizando, assim, as diferenças regionais e conhecendo a vocação do local onde a escola se encontra.

BIBLIOGRAFIA

BERRY, Michael; LINOFF, Gordon. **Data mining techniques**. New York: Wiley, 1997.

BRAGA, Luis Paulo Vieira. **Introdução à mineração de dados**. Rio de Janeiro: E-Papers Serviços Editoriais, 2004.

BRASIL, DEP. **Normas de planejamento e conduta do ensino.** Rio de Janeiro, 2005.

BRASIL, **Manual para avaliação da gestão pública.** Programa da qualidade no serviço público serviço público, 2003.

CASSANO, Daniel. As perspectivas para o mercado de software de gestão. **Falando de Qualidade.** n. 141, fev 2004.

CATTEL, R.B. The scree test fortune number of factors. **Multivariate Behavioral Research**, 1, p. 245-276, 1966.

Clementine **User guide, a data mining toolkit**. Disponível em <http://www.spss.com/clementine/>. Acesso em 2000.

CORNESKY, Robert. **The quality professor: i**mplementing TQM in the classroom. Madison, EUA: Magma Publications, 1993.

DACHS, Norberto. **Estatística computacional**. Rio de Janeiro: Livros Técnicos e Científicos, 1998.

DEMO, P.; RAMOS, C.. **Educação e qualidade:** duas visões, duas orientações. Rio de Janeiro, 1995.

DRUCKER, Peter F., **Knowledge work:** executive excelence. Provo: APR, 2000.

FALCONI, Vicente. **TQC – Controle da Qualidade Total:** no estilo japonês. São Paulo: Fundação Cristiano Ottoni, 2001.

FERRAUDO, Antônio. **Análise multivariada**. São Paulo: StatSoft South América, 2005.

GÓIS, Maurício. Seja um vencedor diante das mudanças. **Banas Qualidade.** n. 111, p. 59, ago 2001.

GIL, Antônio de Loureiro. **Qualidade Total nas Organizações**. São Paulo: Atlas, 1992.

HAYKIN, Simon. **Redes neurais**: princípios e práticas. Porto Alegre: Bookman, 2001.

INMOM, W. H. **Building the data warehouse**. Nova Iorque , EUA: Wiley, 1993.

INMOM, W. H. **The data warehouse and data mining**. Nova Iorque: Wiley. 1996.

JACKSON, J.E. Quality control methods for two related variables. **Industrial Quality Control**, January. p. 4 – 8, 1956.

____ . Principal components and factor analysis: Part I – principal components. **Journal of Quality Technology**, October. v.12, n.4, p.201 – 213, 1981.

____ . Principal components and factor analysis: Part II – additional topics related to principal components. **Journal of Quality Technology**, January, v.13. n.1, p.46 – 58, 1981.

____ . Principal components and factor analysis: Part III – what is factor analysis? **Journal of Quality Technology**, April. v.13, n.2, p.125 – 130, 1981.

JOHANNPETER, Jorge Gerdau. Competitividade e produtividade: Fatores decisivos para a excelência gerencial. **Falando de Qualidade**. n. 139, p. 46. dez 2003.

JOHNSON, Richard A.; WICHERN, Dean W. **Applied multivariate statistical analysis**. Englewood Cliffs, EUA: Prentice Hall, 1992.

KHATTREE, Ravindra; NAIK, Dayanand N. **Multivariate data reduction an discrimination**. Cary, EUA: Wiley Inter-Science, 2001.

HOTTELLING, H. Analysis of a complex of statistical variables into principal components. **The Journal of Educational Psychology**, v.24, p.417 – 441, 498 – 520, 1933.

KLOSGEN, Willi; ZYTKOW, Jan. **Handbook of data mining and knowledge discovery**. Oxford: Oxford University Press, 2002.

LÍRIO, Gilvete S. **Métodos multivariados:** uma metodologia para avaliar a satisfação dos clientes da RBS-TV na região noroeste do RS. 2004. 16f. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria, Santa Maria, 2004.

LOUZADA NETO, F.; DINIZ, C.A.R. **Data mining:** uma introdução. São Paulo: Associação Brasileira de Estatística, 2000.

LOPES. M. P. D. **Gerenciamento da qualidade no ensino da matemática**. Santa Maria: UFSM, 2004. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria, 2004.

MAGNUSSON, Wiliam E.; MOURÃO, Guilherme. **Estatística sem matemática**. Londrina, PR: Planta, 2003.

MALHOTRA, Naresh K. **Pesquisa de Marketing**: uma orientação aplicada. Porto Alegre: Bookman, 2001.

MANLY, B. F. J. **Multivariate statistical methods:** a primer. London: Chapman and Hall, 1986.

MORRISON, D.F. **Multivariate statistical methods.** 2. Ed., New York: Mc Graw Hill, 1976.

PEARSON, K. On lines and planes of closed fit to system of point in space. **Phil. Mag.**, v. 6, p. 559 – 572.

PEREIRA, Júlio C. R. **Análise de dados qualitativos:** estratégias metodológicas para as ciências da saúde, humanas e sociais. São Paulo: Editora da Universidade de São Paulo, 2001.

PEREIRA, Maria José Lara de Bretãs; FONSECA, João Gabriel Marques. **Faces da decisão:** as mudanças de paradigmas e o poder da decisão. São Paulo: Makron Books, 1997.

RAMOS, C. **Sala de aula da qualidade total.** Rio de Janeiro: Qualitymarrk, 1992.

REGAZZI, A. J. **Apostila da disciplina: INF 766:** análise multivariada**.** Viçosa: Universidade Federal de Viçosa. Centro de Ciências Exatas e Tecnológicas. Departamento de Informática, 2001. v.2.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDERSHAN, S. **Sistema de banco de dados**. São Paulo: Makron Books, 1999.

SNEATH, P. H. A.; SOKAL, R. R. **Numerical taxonomy**. San Francisco, USA: Freeman Co.,1973.

SOUZA, A. M. **Componentes Principais**: aplicação na redução de variáveis econômicas para o estudo de séries temporais. Santa Maria: UFSM, 1993. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria, 1993.

_____. **Monitoração e ajuste de realimentação em processos produtivos multivariados**. Florianópolis: UFSC, 2000. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, 2000.

SOUZA, A. M.; SAMOHYL, R. W.; MALAVÉ, C. O. **Multivariate feed control:** an application in a productive process. Computers & Industrial Engineering, Vol 46 Issue 4, Jul 2004.

STAIR. R. M. **Princípios de sistemas de informação: u**ma abordagem gerencial. 2. ed. Rio de Janeiro: LTC, 1998.

STATSOFT, INC. **STATISTICA 7.0** Tulsa, Oklahoma: StatSoft Inc., 2005.

TURKEY, J. **Exploratory data analysis mining.** Nova Iorque: MacMillan, 1973.

VALENTIN, J. L. **Ecologia numérica:** uma introdução à análise multivariada de dados ecológicos. Rio de Janeiro: Interciência, 2000.

VIRGILLITO, Salvatore B. **Estatística aplicada**. São Paulo: Alfa-Omega, 2004.

WERKEMA, M. C. C. **As ferramentas da qualidade no gerenciamento de processos.** Belo Horizonte: Fundação Christiano Ottoni, 1995.