Perfil dos alunos e dos Colégios Militares: um enfoque multivariado

Adriano Mendonça Souza[1], Fernando Monteiro Silva[2]

[1]Departamento de Estatística/CCNE/UFSM
Bolsista CAPES BEX 1784/09-9 - Santa Maria, RS
[2]PPG em Estatística e Modelagem Quantitativa/CCNE/UFSM - Santa Maria, RS
e-mail: amsouza@smail.ufsm.br

Resumo

Com o objetivo de determinar o perfil dos alunos e dos Colégios Militares brasileiros aplicam-se técnicas estatísticas multivariadas em dados de rendimento dos alunos. As técnicas aplicadas em indicadores de ensino mostram-se adequadas para a verificação da qualidade, pois, obedecendo à natureza multivariada, extraem-se informações relevantes, utilizando-se diferentes casos e variáveis. Desta forma, busca-se aumentar a competência e a criatividade nas instituições públicas, visando à organização e à 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, colégio militar, estatística multivariada, análise discriminante.

Abstract

In order to determine the students‘ and the Militar Schools‘ profile, multivariate statistical techniques on the students’ data are applied, providing therefore, a source of information to the administration decision taker. The techniques applied in teaching indicators perform a suitable check to the quality because when the multivariate nature is obeyed, relevant information can be taken by using different variants and cases. Thus, it is aimed to enhance the competence and creativity in public institutions looking for organization and quality system management by the use of quantitative methodologies in order to show the comparative performance among schools and their own pupils.
Key words: teaching, military, multivariate statistic, discriminant analysis

1. Introdução

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

Os colégios podem tratar melhor os desafios relativos ao ensino através da busca de conhecimento em bases de dados estruturadas. Portanto o objetivo desta pesquisa é determinar o perfil dos alunos e dos Colégios Militares, utilizando técnicas estatísticas multivariadas aplicadas ao rendimento escolar dos alunos de forma a estudar a relação entre as variáveis em estudo e detalhar as técnicas estatísticas aplicadas na exploração de dados.

Esta pesquisa torna-se importante na medida em que há necessidade de se conhecer tanto o perfil da escola com 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.

No âmbito administrativo do ensino, através das ferramentas e técnicas estatísticas, disponibilizam-se informações e definem-se padrões, os quais podem auxiliar na elaboração de material didático e ser útil para pesquisas futuras (CORNESKY, 1993).

2. Metodologia

Utilizam-se variáveis numéricas referentes aos graus alcançados pelos alunos nas disciplinas da terceira série do ensino médio de 2004 e ao 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); Matemática (Mat); Língua Portuguesa (Port) e Química (Qui).

Para a análise descritiva (ADE), utilizam-se as variáveis: média geral da série (MGS) e tipo do amparo (TA), em que 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 em relação ao rendimento e à origem dos alunos.

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 Sistema Colégio Militar do Brasil (SCMB), a média de aprovação igual ou superior a cinco.

O TA é uma variável categórica, podendo ser Amparado, Concursado ou Transferido. Os 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.

Os procedimentos multivariados utilizados foram a análise de agrupamentos (AA) e a análise de componentes principais (ACP). Tais técnicas possibilitam 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 serem estudados e investigados os planos fatoriais, a análise discriminante (ADI) será utilizada para identificar, primeiramente, as variáveis que discriminam um aluno que estará na condição ser aprovado, aprovado com recuperação e reprovado, buscando-se, desta forma, verificar se um determinado aluno necessita de auxílio pedagógico para continuar acompanhando o desempenho da turma em que ele está inserido.

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); Matemática (Mat); Língua Portuguesa (Port) e 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.

Por se tratar de um método de classificação de casos, a ADI procura determinar quais disciplinas são mais importantes para a questão da aprovação final.

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. Assim, pode-se identificar em qual classe se enquadra o suposto aluno testado no modelo de 2004.

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

3. Técnicas multivariadas

A utilização de AM é aplicada a diversas áreas do conhecimento, tornando as interpretações mais valiosas quando existe a interação

entre as variáveis. Descrevem-se, a seguir, as técnicas de análise de agrupamentos, componentes principais e discriminante. Dentre os autores consultados, deve-se lembrar JACKSON (1956, 1981) KHATTREE e NAIK (2001), HOTELLING (1933), MAGNUSSON e MOURÃO (2003) e PEREIRA (2001).

Análise de Agrupamento

Um cluster visa agrupar variáveis com características comuns, sem perder informações de todo o conjunto em estudo. A análise de agrupamentos (AA) pode ser utilizada tanto para agrupar elementos amostrais (objetos) quanto para variáveis.

O primeiro passo consiste em formular o problema de aglomeração, definindo as variáveis sobre as quais se baseará a aglomeração. Ao medir a similaridade, a distância euclidiana é uma das mais utilizadas pelos pesquisadores, ressaltando que distâncias menores indicam maior similaridade (PEREIRA, 2004). Mas existem outros cálculos de distância, como a distância generalizada ou ponderada, distância de Minkowsky, distância Euclidiana Média, distância City-Block, distância de Mahalanobis, entre outros métodos de definir a matriz da parecença, amplamente discutidos em Mingoti (2005), Hair *et al.* (2005), Malhotra (2001), bastando o pesquisador decidir pela que melhor se adapte ao seu estudo.

Determinada a matriz de parecença, é necessário escolher o método de aglomeração, pois, como citado anteriormente, todos os objetos começam sozinhos e paulatinamente vão formando-se os clusters. Existem várias maneiras de aglutinar os objetos, as mais populares são o método de ligação simples ou *single linkage*, em que os objetos são aglutinados a um dado nível de distância, se um dos objetos em um dos grupos está àquela distância ou mais próximo de pelo menos um objeto do segundo grupo. A ligação do vizinho mais distante ou *complete linkage* considera que dois grupos devem se unir se e somente se os membros mais distantes dos dois grupos estão próximos o suficiente. A ligação média ou *average linkage* acontece se a distância média entre os dois grupos é pequena o suficiente. Existem outros métodos, como o método de *Ward* e o do centróide, que também são muito utilizados (MANLY, 2008).

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, a distância euclidiana é a mais utilizada.

Análise de componentes principais

Em 1901, Karl Pearson apresentou o método destinado inicialmente ao ajuste de planos em geometria espacial. O objetivo da análise de componentes principais (ACP) é encontrar linhas e planos que melhor se ajustem a um conjunto de pontos em um espaço p-dimensional.

Primeiramente, é necessário calcular a matriz de variância-covariância ($\sum$)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).

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.

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 $\hat{\ell}_1'\hat{\ell}_1 = 1$, onde os coeficientes encontrados devem satisfazer as *p-equações* lineares simultaneamente (HAIR et al, 2005; MALHOTRA, 2001).

A variância amostral da *j-ésima* componente é $\hat{\Lambda}_j$, e a variância total do sistema é: $\hat{\Lambda}_1 + \cdots + \hat{\Lambda}_p = tr\,S$. O grau de explicação fornecido pela *j-ésima* componente é fornecida por: $\dfrac{\hat{\Lambda}_j}{tr\,S}$.

A definição do número de componentes a serem utilizadas é feita por meio do método gráfico devido a Catell (1966) e pelo método do Kaiser (1960) *apud* Souza (2000), que consiste em incluir somente aquelas componentes cujos valores próprios sejam superiores a 1, e, em geral, utilizam-se aquelas componentes que conseguem sintetizar uma variância acumulada em torno de 70% (SOUZA, 2000; JOHNSON e WICHERN, 1992). A escolha do número de componentes usando o valor 1 como limite inferior nem sempre é a mais adequada assim como não há de fato um valor de referência que seja genericamente aceito como o mínimo estabelecido para a escolha do número de componentes.

Análise Discriminante

A análise discriminante (ADI) é a técnica de dependência multivariada aplicada quando a variável dependente é qualitativa e possui duas ou mais características, e as variáveis independentes são quantitativas, havendo necessidade de identificar a qual dos grupos pertence um certo caso (indivíduo ou observação) (VIRGILLITO, 2004). Segundo Malhotra (2001), os objetivos da ADI são: estabelecer funções discriminantes que melhor discriminem entre as categorias da variável dependente (grupos); verificar se existem diferenças significativas entre os grupos, em termos das variáveis prognosticadoras; determinar as variáveis preditoras que mais contribuam para as diferenças entre grupos; enquadrar ou classificar os casos em um dos grupos, com base nos valores das variáveis preditoras; avaliar a precisão da classificação.

Segundo Ferraudo (2005), a ADI implica obter um valor teórico, que é uma combinação linear das variáveis independentes que discrimine melhor entre os grupos definidos, *a priori*, segundo:

$$Z_{jk} = a + W_1X_{1k} + W_2X_2 + W_3X_{3k} + ...+ W_nX_{nk} \quad (1)$$

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:

$$Z_{CE} = \frac{Z_A + Z_B}{2} \quad (2)$$

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:

$$Z_{CD} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} \qquad (3)$$

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 $Z_n < Z_{CT}$ e 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 do 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, μ_2 e Σ, respectivamente, os vetores médios de e de , e a matriz de covariância comum a ambas as populações, tem-se a função a seguir:

$$y = (\mu_1 - \mu_2)'\Sigma^{-1} X \qquad (4)$$

Então, pode-se expressar a regra de classificação para X_0, como: alocar X_0 em Φ_1 se y_0 - m e" 0 ou alocar X_0 em Φ_2 se y_0 - m < 0.

Na realidade, os parâmetros μ_1, 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:

$$y = (X_1 \quad X_2)' S_p^{\ 1} X \qquad (5)$$

O valor de corte m é estimado por:

$$m = \frac{1}{2}(y_1 + y_2) \qquad (6)$$

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

$$(X_1 - X_2)' S_p^{-1} X \geq \text{m}, \qquad (7)$$

ou alocar X_0 em G_2 se

$$(X_1 \quad X_2)' S_p^{\ 1} X \quad < \text{m} \qquad (8)$$

Há três razões principais para se desenvolver um projeto de exploração de dados, que são: a visualização dos dados, a descoberta de novos conhecimentos e a acuracidade dos dados.

4. Resultados

Atualmente, o Sistema Colégio Militar do Brasil (SCMB) de escolas militares é formado por 12 (doze) Colégios Militares, situados em 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.

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

Análise descritiva

Para traçar o perfil dos alunos e dos Colégios em estudo, inicialmente aplica-se uma análise descritiva, cuja a população em estudo é composta por 3360 alunos dos quatro Colégios Militares, com a seguinte configuração: Colégio Militar de Belo Horizonte – CMBH, 695 alunos; Colégio Militar de Curitiba - CMC, 725 alunos; Colégio Militar do Rio de Janeiro – CMRJ, 1276 alunos e Colégio Militar de Santa Maria – CMSM com 664 alunos.

Nesta análise, procura-se relacionar o rendimento com a origem do aluno, em que o rendimento é representado pela variável Média Geral da Série (MGS), dos quais 1798 são amparados, 1507 concursados e 55 transferidos, totalizando 3360 alunos.

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.

Os Colégios Militares, de modo geral, apresentam uma concentração maior de alunos concursados com rendimento Bom e Muito bom, enquanto os alunos amparados concentram-se no rendimento bom. Ainda, nota-se que o rendimento abaixo da média cinco, ou seja, com menção Insuficiente, encontra-se apenas nos alunos amparados.

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.

Nota-se uma baixa proporção de alunos amparados 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 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.

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

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, porque apenas esses colégios 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 notas finais das disciplinas de 184 alunos da 3ª série do Ensino Médio, no mesmo ano letivo da análise anterior.

Os valores das médias e desvios padrão das variáveis, utilizando todos os casos deste estudo, estão representados na Tabela 1.

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.

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

Nesse caso, utiliza-se o processo de aglomeração hierárquica,

com o método de variância de Ward e distância euclidiana.

A Figura 1 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.

Tabela 1. Médias e desvio padrão das variáveis.

Variáveis	Média	Desvio Padrão	Mínimo	Máximo
GrauComp	9,433	1,165	3,400	10,000
Bio	7,137	0,992	5,000	9,200
EF	7,671	1,676	0,900	10,000
Fis	7,229	0,978	5,300	9,800
Geo	7,398	0,751	5,700	9,400
Hist	7,902	1,019	5,100	9,700
LEM	7,293	0,997	4,900	9,500
Lit	6,848	0,917	4,700	9,700
Port	7,125	0,831	5,100	9,700
Mat	6,841	1,304	3,400	9,700
Qui	7,132	0,999	5,200	9,500

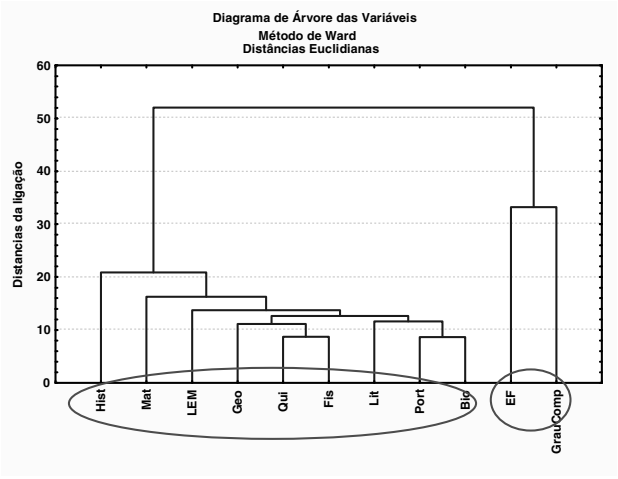

Figura1. Dendograma envolvendo as variáveis em estudo.

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.

As variâncias explicadas por cada autovalor, os autovalores acumulados e suas respectivas variâncias acumuladas, calculados a partir da matriz de correlação, são mostradas na Tabela 2.

Observa-se que esta explicação é devido 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 2.

Tabela 2. Autovalores e percentual de variância explicada.

Fatores	Autovalores	Variância Explicada (%)	Autovalores Acumulados	Variaância 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

O percentual de variância explicada pelos dois primeiros autovalores é de 65,617%, que representa a variabilidade total do sistema.

O critério da escolha do autovalor maior que um e o gráfico da Figura 2 corroboram para serem usadas apenas as duas primeiras componentes para a avaliação das variáveis, 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 4.

Os resultados mostram a relação das componentes principais com as variáveis originais, as quais, na Figura 3 apresenta-se no plano fatorial a relação das componentes principais com as variáveis originais e de que forma estas variáveis se agrupam, confirmando os resultados dos agrupamentos formados e apresentados no dendograma da Figura 1.

Tabela 3. 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

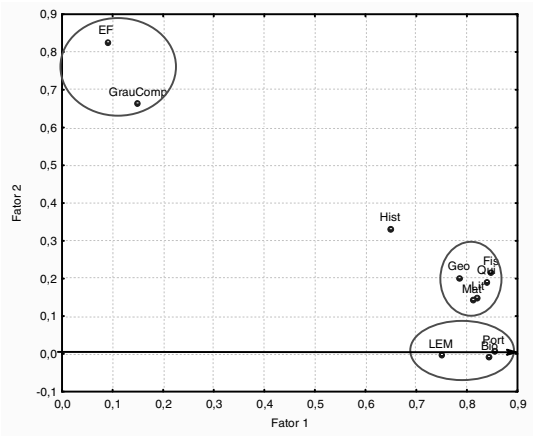

Figura 3. Plano Fatorial – Fator 1 *vs* Fator 2.

Isso não impede a realização de um estudo da correlação entre os fatores e os dados originais para salientar quais são as variáveis mais representativas cada fator.

Deve-se buscar uma interpretação física para melhor entender esses fatores. Depois de definidos os fatores de estudo, representam-se graficamente, na Figura 03 as variáveis no plano fatorial para comprovar os agrupamentos formados.

Nota-se que os agrupamentos são semelhantes aos formados na AC, representando o Fator 1 como área das ciências cognitivas, que exige estudo, escrita e leitura, e o fator dois, os atributos da área psicomotora/ afetiva. Nota-se, aqui, um maior distanciamento de História das demais disciplinas formadoras do agrupamento da área das ciências cognitivas, 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, em que 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.

Devido à comprovação destes grupos formados, foram escolhidos aleatoriamente seis alunos, três de cada Colégio, como forma de verificar a sua relação com os planos fatoriais. Na Figura 4 visualizam-se os alunos no plano fatorial.

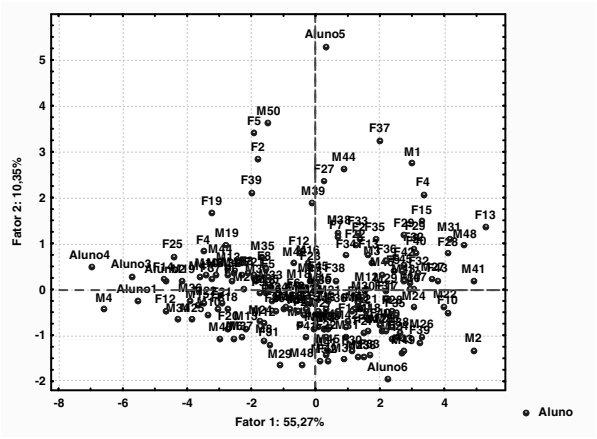

Figura 4. Projeção dos casos no plano fatorial.

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 3.

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, tornando-se válida a análise, pois foi possível caracterizar o perfil desses alunos.

Análise discriminante

Utiliza-se nessa análise dados de rendimentos finais de 582 alunos do CMC e CMSM com resultados do final do ano letivo de 2004 das disciplinas e pontos perdidos, relacionados com a variável categórica Situação da Matrícula (Aprovado, Aprovado com PR e Reprovado).

Para esse estudo, utilizou-se o efetivo do Ensino Médio do CMC e CMSM. A seguir, são coletados dados individuais dos elementos de cada grupo. Por se tratar de um método de classificação de casos, a ADI, procura determinar quais disciplinas são mais importantes para a questão da aprovação final.

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. Assim, pode-se identificar em qual classe se enquadra o suposto aluno testado no modelo de 2004.

Na Tabela 5 é apresentado 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, em que 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 4. 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

O Lambda Wilks é uma estatística que pode ser auxiliar para testar a contribuição das variáveis independentes na função discriminante. Dentre as vaiáveis selecionadas para a função discriminante, observa-se que todas são significativas apresentando um p-valor próximo de zero. Mas a variável "Fis" é a que apresenta maior poder de discriminação CORRAR, PAULO e DIAS FILHO (2007).

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).

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

$$Y_{APROVADOS} = 1{,}14 \cdot Fis + 7{,}94 \cdot Geo + 0{,}73 \cdot Mat - 36{,}86$$
$$Y_{APROVADOS\ C/PR} = 0{,}02 \cdot Fis + 6{,}65 \cdot Geo + 0{,}73 \cdot Mat - 23{,}93$$
$$Y_{REPROVADOS} = 0{,}76 \cdot Fis + 7{,}72 \cdot Geo - 0{,}57 \cdot Mat - 28{,}63$$

A Matriz de Classificação, apresentada na Tabela 5, demonstra o percentual de validação da função discriminante, em que 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 5. 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

Após a identificação das variáveis significantes, parte-se para uma aplicação prática, em que informa o provável grau para as disciplinas selecionadas pela função discriminante e apresenta-se um resultado gerado pela classificação. A Tabela 6 apresenta as médias das variáveis, para cada tipo de situação de matrícula.

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 para Física, o grau igual a 6. Para a classificação

do aluno, foi utilizada a distância de Mahalanobis, descrita no item 2.

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.

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.

Tabela 6. Média das variáveis e situações de matrícula.

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

5. Conclusões e sugestões

Na primeira análise, pode-se identificar um padrão entre os Colégios e classificar as escolas de acordo com o modelo formado, em que 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.

Na segunda análise, verifica-se a relação entre as disciplinas e o comportamento, em que 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.

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 pode-se 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.

Esta pesquisa é importante para os Colégios Militares, pois, utilizando 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.

Objetivando aumentar a 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.

6. Bibliografia

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

CORRAR J.L.; PAULO, E e DIAS FILHO, J.M. **Análise Multivariada: para cursos de administração, ciências contábeis e econmomia.** Ed. Atlas. São Paulo. 2007

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

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

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

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

HAIR, J.F. Jr. et al.; trad. Sant´Ana A. S. e Neto, A.C. **Análise multivariada de dados.** 5 ed. Porto Alegre: Bookman. 2005.

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

KHATTREE, R.; NAIK, D. N. **Multivariate data reduction and discrimination: with SAS software.** SAS institute. John Wiley & Sons, Inc. Cary. North Caroline, USA. 2000.

MAGNUSSON, W. E.; MOURÃO, G. **Estatística sem matemática**. Londrina, PR: Planta, 2003.

MALHOTRA, N. 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. 1901

PEREIRA, J. 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.

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

Submetido em: 15/07/2008
Aceito em: 30/09/2009