UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

Jeferson Nomelini

APLICAÇÃO DE METODOLOGIAS DE EXTRAÇÃO DE CONHECIMENTOS EM PESQUISAS E GERENCIAMENTO DE PROGRAMA DE MELHORAMENTO GENÉTICO EM BOVINOS DA RAÇA NELORE

Ribeirão Preto – SP
2006

JEFERSON NOMELINI

APLICAÇÃO DE METODOLOGIAS DE EXTRAÇÃO DE CONHECIMENTOS EM PESQUISAS E GERENCIAMENTO DE PROGRAMA DE MELHORAMENTO GENÉTICO EM BOVINOS DA RAÇA NELORE

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo como requisito parcial para obtenção do título de Mestre em Ciências.
Área de concentração: Genética

Orientador: Prof. Dr. Raysildo Barbosa Lôbo

Ribeirão Preto – SP
2006

FICHA CATALOGRÁFICA

Nomelini, Jeferson

Aplicação de Metodologias de Extração de Conhecimentos em Pesquisas e Gerenciamento de Programa de Melhoramento Genético em Bovinos da Raça Nelore / Jeferson Nomelini; Orientador Prof. Dr. Raysildo Barbosa Lôbo. Ribeirão Preto, 2006.

113 f.: fig.

Dissertação (Mestrado – Programa de Pós-graduação em Genética), – Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.

Orientador: Lôbo, Raysildo B.

1- Nelore, 2- melhoramento genético, 3- *data warehouse,* 4- *OLAP*, 5- mineração visual de dados, 6- progresso genético.

FOLHA DE APROVAÇÃO

Jeferson Nomelini

Aplicação de Metodologias de Extração de Conhecimentos em Pesquisas e Gerenciamento de Programa de Melhoramento Genético em Bovinos da Raça Nelore.

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo como requisito parcial para obtenção do título de Mestre em Ciências.
Área de concentração: Genética

Aprovado em: ____/____/____

Banca Examinadora

Prof. Dr. _____

Instituição:_____Assinatura;_____

Prof. Dr. _____

Instituição:_____Assinatura;_____

Prof. Dr. _____

Instituição:_____Assinatura;_____

Dedico este trabalho aos meus pais,
Lourenço e Zeli, meu irmão, Robson,
e meus sobrinhos Rafael e Gustavo.

AGRADECIMENTOS

À **Deus**, pela vida, sabedoria e fé.

Aos meus **pais** e **família**, por sempre me apoiarem nos estudos.

Ao **Prof. Dr. Raysildo Barbosa Lôbo**, pela amizade, orientação e oportunidades.

À **Profa. Dra. Solange Oliveira Rezende**, pela valiosa "co-orientação".

Aos amigos **Rita de Cássia Pereira Lôbo**, **Pedro Alejandro Vozzi** e **Milena Cereli**, pelas revisões de meu trabalho.

Ao amigo **Valmir Marques Ferreira**, pelos ensinamentos e manutenção do *Nelore Business Intelligence*.

Aos técnicos **Luiz Bezerra**, **Paulo Bezerra** e **Marco Corrado** pelo aporte em informática e amizade.

Aos amigos **Cíntia Righetti Marcondes** e **José Eduardo do Val** pelas calorosas discussões sobre o melhoramento genético e proveitosas críticas aos meus trabalhos.

Aos **professores**, **funcionários** e **alunos** do **Departamento de Genética da FMRP-USP**, pela amizade, convivência e trocas de experiência.

Aos **Pesquisadores do IZ de Ribeirão Preto**, pela colaboração no ensino e pesquisa da Genética Quantitativa junto ao GEMAC.

Aos **funcionários e colaboradores da ANCP**, pela amizade e o excelente trabalho prestado.

À **USP**, pelo ensino de excelência que me proporcionou.

À **todos meus amigos**, que com certeza, contribuíram para minha vitória.

Às instituições **PRONEX**, **FAPESP**, **CNPq**, **CAPES** e **ANCP**, pelo suporte financeiro essencial a minha pesquisa.

São dados a ti a sabedoria e o conhecimento; também te darei riquezas, e bens materiais, e honra, tais como nenhum rei anterior a ti veio a ter, e tais como nenhum depois de ti virá a ter.

II Crônicas 1:12

RESUMO

O Programa de Melhoramento Genético da Raça Nelore (PMGRN – Nelore Brasil) experimentou, em 17 anos de existência, rápido crescimento geográfico, abrangendo rebanhos em 12 estados brasileiros mais 3 países da América Latina, além de crescimento exponencial de sua base de dados (774.870 animais avaliados em 2005). Para administrar essa enorme base de dados e gerar descobertas científicas que proporcionem progresso genético aos rebanhos avaliados, torna-se necessário o uso de tecnologias da informação. Os objetivos foram utilizar tecnologias da informação para: caracterizar a estrutura populacional do rebanho avaliado, identificar vias de fluxo gênico, analisar evolução do coeficiente de endogamia e extrair padrões de seleção e acasalamento das fazendas participantes. Foi utilizado o *Nelore Business Intelligence*, o sistema de inteligência empresarial (*BIS*) do PMGRN – Nelore Brasil, carregado com a avaliação genética de 2005, como fonte única de dados. As tecnologias utilizadas para análise foram Processamento Analítico *On-Line* (*OLAP*) e mineração visual de dados. Quanto ao sexo dos animais avaliados, geralmente há equilíbrio em regiões exportadoras de genética (rebanho *seleção* – destinados à produção de reprodutores) e maior proporção de matrizes em regiões exportadoras de carne (rebanho *multiplicador* – destinado à multiplicação dos genes provenientes do rebanho *seleção* e rebanho *comercial* – dedicado à produção de carne), além disso, a proporção vem equilibrando ao longo do tempo, indicando amadurecimento do PMGRN – Nelore Brasil. Quanto à variedade (padrão e mocho), o mocho representa 24% de todo rebanho avaliado, com ligeiro crescimento de participação nos últimos anos, preferência e crescimento regional de cada uma das variedades provavelmente estejam ligadas às culturas e tradições locais e marketing dos criadores. O fluxo gênico na Raça Nelore ocorre pela incorporação de genes do rebanho *seleção* no *multiplicador* e *comercial*, causado, principalmente, pela difusão da Inseminação Artificial, proporcionando progresso genético até no rebanho *comercial*. O controle da endogamia mostrou-se eficiente para todas categorias, exceto *Puro de Origem Importado*, causado pelo pequeno número de progenitores. A análise da endogamia por fazenda indica bom controle deste índice na maioria delas (92,7% de rebanhos avaliados com coeficiente de endogamia menor que 2%). Animais de elevado potencial genético para diversas características (peso, habilidade maternal, fertilidade e reprodutivas) tem maior probabilidade de produzir grande número de descentes, traduzindo em vantagem competitiva às fazendas que produzem este perfil de reprodutores. A maioria das fazendas participantes do PMGRN – Nelore Brasil selecionaram matrizes com elevado potencial genético para serem destinadas às biotecnologias de Transferência de Embriões e Fertilização *In Vitro*. O uso de um *BIS* é potencialmente eficiente como ferramenta gerencial e fonte de pesquisa de um programa de melhoramento genético, por capacitarem tecnologias *OLAP* e mineração visual de dados a extraírem informações e conhecimentos válidos e úteis para os rebanhos participantes. A aplicação destas tecnologias e a incorporação dos conhecimentos à comunidade de usuários (criadores) podem maximizar o progresso genético do rebanho e aumentar a competitividade da pecuária de corte brasileira.

Palavras chaves: 1- Nelore, 2- melhoramento genético, 3- *data warehouse*, 4- *OLAP*, 5- mineração visual de dados, 6- progresso genético.

ABSTRACT

In the 17 years since its introduction, the Program for Genetic Improvement of the Nellore Breed (PMGRN – Brazilian Nellore) has experienced a rapid expansion. At present, it comprises herds in 12 Brazilian states and in 3 other Latin American countries, and its database has grown exponentially (774,870 assessed animals in 2005). In order to handle such a huge database and produce new scientific knowledge that may lead to the genetic improvement of the assessed herds, the use of information technology is required. The goals of this study were to use IT in order to: determine the population structure of the assessed herd, identify gene flow paths, study the dynamics of endogamy coefficient, and identify selection and mating patterns on the participating farms. *Nellore Business Intelligence*, the business intelligence system (BIS) of PMGRN – Brazilian Nellore, was used and its sole source of input was the genetic assessment of 2005. The analysis was carried out through On-Line Analytical Processing (OLAP) and visual data mining. With regard to the gender of the assessed animals, there is an overall balance among genetic exporting regions (selection herd – aimed at producing breeders) and a larger ratio of dams in beef exporting regions (multiplication herd – aimed at replicating genes that originate from the selection herd and the commercial herd – aimed at producing beef). In addition, that ratio has become more balanced over the years, which suggests the PMGRN – Brazilian Nellore program has been continually honed. As for strains (Nellore and Polled Nellore), Polled Nellore individuals account for 24% of the entire assessed herd, with a slight increase in participation over the last years. Preference and regional growth for each strain may be linked to local cultures, traditions and the marketing strategies used by ranchers. The Nellore gene flow occurs due to the introduction of genes from the selection herd into the multiplication and commercial herds, which is particularly the result of the dissemination of artificial insemination, which leads to the genetic improvement even in the commercial herd. Inbreeding control has proven an efficient strategy for all strains, except Pure of Imported Origin, due to the small number of parents. Inbreeding analysis for each farm reveals there is good control of this index in most of them (92.7% of the assessed herds yields lower than 2% endogamy coefficient). Animals with a high genetic potential for several traits (weight; mothering, fertility and reproductive abilities) are more likely to produce a large number of descendants and, therefore, add competitive advantages to the farms that raise animals of their kind. Most farms in the PMGRN – Brazilian Nellore select dams with high genetic potential so they can be used for embryo transfer and in-vitro fertilization biotechnology procedures. The use of a BIS is potentially efficient as a management tool and as a research source in a genetic improvement program because it allows for the use of OLAP technologies and visual data mining to produce information that can be useful and valid for the participating herds. The use of such technologies and the addition of this knowledge to the community of users (ranchers) may maximize genetic improvement of herds and enhance competitiveness for Brazilian beef cattle ranching.

Key Words: 1- Nellore, 2- genetic improvement, 3- data warehouse, 4- OLAP, 5- visual data mining, 6- genetic progress.

SUMÁRIO

1. INTRODUÇÃO..... 19

1.1. Pecuária brasileira..... 20

1.2. Princípios do melhoramento genético..... 22

1.3. PMGRN – Nelore Brasil..... 25

1.4. Motivação e relevância do uso da tecnologia da informação pelo PMGRN – Nelore Brasil..... 27

2. OBJETIVOS..... 30

2.1. Objetivos gerais..... 31

2.2. Objetivos específicos..... 31

3. REVISÃO BIBLIOGRÁFICA..... 32

3.1. Princípios da tecnologia da informação..... 33

3.2. *Data Warehouse*..... 35

3.3. Consulta *OLAP*..... 40

3.3.1. Facilidade de uso..... 41

3.3.2. Segurança..... 41

3.3.3. Quatro tipos diferentes de relatórios..... 42

3.3.4. Liberdade de seleção da área empresarial e objetos internos..... 43

3.3.5. Aplicação de filtros..... 44

3.3.6. Cálculos..... 45

3.3.7. Funções OLAP..... 46

3.4. Mineração de dados..... 47

3.5. Visualização de dados..... 55

3.5.1. Técnicas de visualização usuais..... 63

3.5.2. Técnica de visualização multidimensional orientada a pixel..... 63

3.5.3.Técnica de visualização multidimensional de projeção geométrica..... 64

4. MATERIAIS E MÉTODOS..... 67

5. RESULTADOS E DISCUSSÃO..... 73
5.1. Caracterização da estrutura populacional da Raça Nelore..... 74
5.2. Fluxo gênico na Raça Nelore..... 83
5.3. Coeficiente de endogamia por categoria animal, safra e fazenda..... 88
5.4. Identificação de padrões de seleção e acasalamento da Raça Nelore..... 91
6. CONCLUSÕES E PROPOSTOS FUTURAS..... 103
6.1. Conclusões..... 104
6.2. Propostas futuras..... 105

REFERÊNCIAS..... 106

LISTA DE FIGURAS

Figura 1 – Contextualização..... 20

Figura 2 – PIB brasileiro e representatividade do agronegócio..... 21

Figura 3 – Distribuição de uma característica hipotética em duas gerações, paterna (*P*) e filial (*F*), de um rebanho sob seleção, assinalando a média da geração paterna ($\overline{X}_p$), a média dos indivíduos selecionados ($\overline{X}_s$), a média da geração F ($\overline{X}_f$) e a resposta à seleção (*R*)..... 24

Figura 4 – Distribuição dos animais avaliados pelo PMGRN – Nelore Brasil em função da área geográfica (estado brasileiro ou país da América Latina) e porcentagem de animais nascidos em função do período compreendido..... 26

Figura 5 – Metodologia: dos dados à vantagem competitiva..... 28

Figura 6 – Arquitetura do *Nelore Business Intelligence*..... 37

Figura 7 – Esquema constelação do *PMGRN-DM*..... 38

Figura 8 – Menu de conexão do *Discoverer* ao *Data Warehouse*..... 42

Figura 9 – Menu com tipos de relatórios..... 43

Figura 10 – Menu de seleção da área de trabalho, grupo de objetos e objetos..... 44

Figura 11 – Menu para aplicação de filtros..... 44

Figura 12 – Menu para editar cálculos..... 45

Figura 13 – Funções *OLAP*..... 47

Figura 14 – Relações entre base de dados e *Data Warehouse*; diferenças entre os processos de *Data Warehousing* (*DW*), consulta *OLAP* e mineração de dados (*DM*); natureza do conhecimento humano..... 49

Figura 15 – Etapas e usuários do processo de mineração de dados..... 50

Figura 16 – Arquitetura do *Spotfire*..... 57

Figura 17 – Tela do *Spotfire* e seus componentes..... 58

Figura 18 – Importação de dados de um *SGBD Oracle* via *OLE DB*..... 60

Figura 19 – Seleção de atributos do *Data Warehouse* para visualização no *Spotfire*..... 60

Figura 20 – Exemplo de codificação numérico-categórica..... 61

Figura 21 – Exemplo de construção de atributo..... 62

Figura 22 – Exemplo de menu Propriedades..... 62

Figura 23 – Técnica de visualização multidimensional orientada a *pixel*..... 64

Figura 24 – Técnica de visualização multidimensional de coordenadas paralelas: Mapa de Perfil..... 66

Figura 25 – Técnica de visualização multidimensional de dispersão: A) Gráfico de Dispersão em duas dimensões; B) Gráfico de Dispersão em três dimensões; C) Gráfico de Matrizes..... 66
Figura 26 – Distribuição dos animais por sexo..... 75
Figura 27 – Distribuição dos animais por sexo e categoria..... 75
Figura 28 – Distribuição dos animais por sexo, estados brasileiros e outros países..... 76
Figura 29 – Distribuição dos animais por sexo e qüinqüênios..... 76
Figura 30 – Distribuição dos animais por categoria..... 77
Figura 31 – Distribuição dos animais por classes de categorias, estados e outros países.... 78
Figura 32 – Distribuição dos animais por categoria e qüinqüênios..... 78
Figura 33 – Distribuição dos animais por variedade..... 80
Figura 34 – Distribuição dos animais por variedade da progênie e progenitores..... 80
Figura 35 – Distribuição dos animais por variedade e categoria..... 81
Figura 36 – Distribuição dos animais por variedade, estados brasileiros e outros países e qüinqüênios..... 81
Figura 37 – Crescimento da participação das variedades por qüinqüênios (valores positivos no gráfico indicam crescimento da participação da variedade mocha e valores negativos, da variedade padrão)..... 82
Figura 38 – Fluxo gênico na pecuária de corte..... 83
Figura 39 – A) Progresso genético para MGT por classe de categorias; B) Estatística descritiva..... 85
Figura 40 – A) Distribuição do número de animais concebidos pelos acasalamentos entre grupos de categorias dos progenitores; B) Investigação de Detalhes para categoria da progênie em cada grupo de acasalamentos..... 86
Figura 41 – Tipo de acasalamento por qüinqüênios..... 87
Figura 42 – Evolução da endogamia por categoria animal..... 89
Figura 43 – Relação progênies e progenitores por categoria..... 89
Figura 44 – Fazendas com maiores médias de endogamias para a safra 2003..... 90
Figura 45 – A) Progresso genético do MGT comparativo do PMGRN – Nelore Brasil com as fazendas que apresentaram média de F ≥ 2,0% na safra 2003; B) Estatística descritiva..... 90
Figura 46 – Distribuição dos animais nascidos entre 1994 e 2003 por estado brasileiro e outros países: A) Distribuição de todos animais; B) Distribuição dos animais TOP 25%; C) Porcentagem dos animais TOP 25%..... 94
Figura 47 – Distribuição dos animais por tipo de acasalamento e classe..... 95
Figura 48 – Distribuição dos animais TOP 25% por categoria..... 95

Figura 49 – Relações entre animais TOP 25%, MGT e fazenda..... 96
Figura 50 – Perfil genético dos animais TOP 25%..... 96
Figura 51 – Relações dos animais TOP 25% com número de filhos que deixaram avaliados no rebanho (NF120)..... 97
Figura 52 – Relações das matrizes BOTTON 50% que foram submetidas a TE ou FIV com número de filhos avaliados aos 120 dias de idade (NF120) e tipo de acasalamento (TA) que deram origem a essas matrizes..... 100
Figura 53 – Relações das progênies oriundas das matrizes BOTTON 50% com o tipo de acasalamento que originaram as progênies, por safra..... 101
Figura 54 – Relações das progênies oriundas das matrizes BOTTON 50% que foram submetidas a TE ou FIV com a matriz e a fazenda onde ocorreu o parto..... 101
Figura 55 – Perfil genético das progênies das matrizes BOTTON 50% que foram submetidas a TE ou FIV..... 102

LISTA DE TABELAS

Tabela 1 – Importância da carne bovina na alimentação..... 22
Tabela 2 – Situação hipotética de três touros para comparação de DEPs..... 34
Tabela 3 – Diferenças entre sistema operacional e sistema de inteligência empresarial (*BIS*) quanto às expectativas do usuário..... 35
Tabela 4 – Relação entre as características do *Data Warehouse* e exemplo no *Nelore Business Intelligence*..... 36
Tabela 5 – Dimensões, atributos e hierarquias do *PMGRN-DM*..... 39
Tabela 6 – Classificação dos níveis de mineração de dados quanto à interação com o especialista do domínio..... 55
Tabela 7 – Atributos e operações utilizadas para análise do progresso genético do MGT. 70
Tabela 8 – Atributos e operações usados para análise do coeficiente de endogamia (F).... 70
Tabela 9 – Codificação numérico-categórica das DEPs..... 71
Tabela 10 – Coeficiente de correlação de Pearson para os conjuntos de dados: todos os animais (linha superior) e TOP 25% (linha inferior)..... 93
Tabela 11 – Relação entre número de filhos avaliados e classe dos touros..... 98
Tabela 12 – Relação entre número de filhos avaliados e classe das matrizes..... 98
Tabela 13 – Valores máximos, mínimos e medianas das DEPs de animais provenientes de matrizes BOTTON 50% que foram submetidas a TE ou FIV..... 102

LISTA DE ABREVIATURAS E SIGLAS

A – Efeito aditivo dos genes

ABCZ – Associação Brasileira dos Criadores de Zebu

ANCP – Associação Nacional de Criadores e Pesquisadores

AVG – Média aritmética

BIS – *Business Intelligence System* ou Sistema de Inteligência Empresarial

BLUP – *Best Linear Unbiased Prediction* ou Melhor Preditor Linear não Viesado

BA – Bahia

BO – Bolívia

BOTTON – Parte inferior

CASE – *Computer-Aided Software Engineering* ou Ferramenta de Engenharia de Software

CEPEA – Centros de Estudos Avançados em Economia Aplicada

CL – Cara Limpa

Count – Operação matemática de contagem

D – Efeito de dominância dos genes

DEP – Diferença Esperada na Progênie

DF – Distrito Federal

DG – Departamento de Genética

DIPP – DEP para efeito direto da idade ao primeiro parto

DP120 – DEP para efeito direto do peso aos 120 dias

DP365 – DEP para efeito direto do peso aos 365 dias

DP450 – DEP para efeito direto do peso aos 450 dias

DPAC – DEP para efeito direto da produtividade acumulada

DPE365 – DEP para efeito direto do perímetro escrotal aos 365 dias

DPE450 – DEP para efeito direto do perímetro escrotal aos 450 dias

E – Efeito ambiental

ETL – *Extraction, Transformation and Load* ou Extração, Transformação e Povoamento

Ex. – Exemplo

F – Coeficiente de endogamia

FIV – Fertilização *in Vitro*

FMRP – Faculdade de Medicina de Ribeirão Preto

GE – Interação genótipo-ambiente

GEMAC – Grupo de Genética, Melhoramento Animal e Computação

GO – Goiás

I – Efeito epistático dos genes

IA – Inseminação Artificial

ICMC – Instituto de Ciências Matemáticas e de Computação

IEA – Instituto de Economia Aplicada

KDD – *Knowledge Discovery in Databases* ou Descoberta de Conhecimento em Base de Dados

LA – Livro Aberto

LABIC – Laboratório de Inteligência Computacional

LT – Lote de touros

MA – Maranhão

Máx – Valor máximo de um conjunto de dados

Mín – Valor mínimo de um conjunto de dados

MG – Minas Gerais

MGT – Mérito Genético Total

MP120 – DEP para efeito maternal no peso aos 120 dias

MS – Mato Grosso do Sul

MT – Mato Grosso

N – Operação matemática de contagem

NF – Número de Filhos

NF120 – Número de progênies com dados válidos para o P120

NR – Número de Rebanhos

NR120 – Número de rebanhos com progênies válidas para o P120

Obs. – Observação

OLAP – *On-Line Analytical Processing* ou Processamento Analítico *On-Line*

OLE DB – *Object Linking and Embedding for Databases* – meio utilizado pela Microsoft para acessar dados armazenados de diferentes formas

OLTP – *On-Line Transactional Processing* ou Processo de Transação *On-Line*

P – Efeito fenotípico

PA – Pará

PE – Perímetro Escrotal

PIB – Produto Interno Bruto

PMGRN – Nelore Brasil – Programa de Melhoramento Genético da Raça Nelore

PMGRN-DM – *Data Mart* referente à avaliação genética

PO – Puro de Origem

POI – Puro de Origem Importado

PR – Paraná

RDBMS – Tecnologia de Banco de Dados Relacionais

RO – Rondônia

ROLAP – Processamento Analítico *On-Line* do tipo Relacional

SGBD – Sistema Gerenciador de Banco de Dados

SIC – Serviço de Informação da Carne

SisNe – Sistema Nelore

SP – São Paulo

TE – Transferência de Embrião

TO – Tocantins

TOP – Parte superior

USA – Última Situação do Animal

USP – Universidade de São Paulo

VE – Venezuela

LISTA DE SÍMBOLOS

%: Porcentagem

µg: Microgramas

cal: Calorias

cm: Centímetros

g: Gramas

Kg: Quilogramas

mg: Miligramas

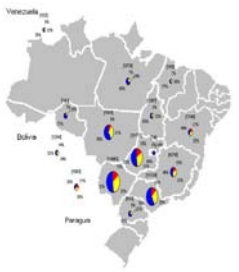
mm: Milímetros

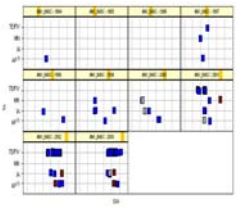
R$: Reais

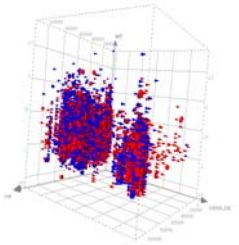
s: Desvio padrão

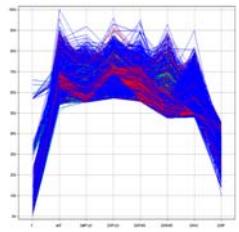
u.d.p.g.: Unidades de desvio padrão genético

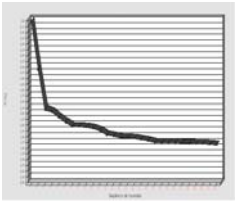
$\bar{x}$ **:** Média aritmética











INTRODUÇÃO

1. INTRODUÇÃO

Os impactos e transformações que a tecnologia da informação vem causando mundialmente nas últimas décadas é semelhante à revolução industrial ocorrida no século XVIII. Enquanto que a revolução industrial marcou um grande salto tecnológico nos transportes e máquinas, a tecnologia da informação vem causando nos processos administrativos das corporações.

Um setor produtivo de destaque na economia brasileira, como a pecuária de corte, necessita de investimentos em tecnologias da informação para ser competitivo frente ao mercado nacional e internacional.

Esta Seção contextualiza o leitor, partindo da visão macro da pecuária brasileira e viajando em diferentes níveis, até a visão micro das contribuições da tecnologia da informação ao Programa de Melhoramento Genético da Raça Nelore (PMGRN – Nelore Brasil) (Figura 1).

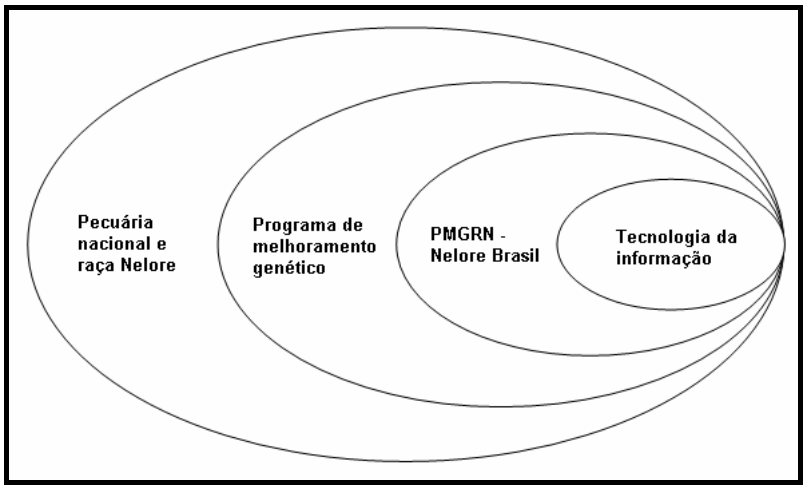


Figura 1 – Contextualização.

1.1. Pecuária brasileira

Dados econômicos (Centros de Estudos Avançados em Economia Aplicada [CEPEA], 2005) demonstram que o agronegócio brasileiro representa, anualmente, uma importante fatia do Produto Interno Bruto (PIB) brasileiro, em torno de 20% de toda a riqueza produzida pelo país no início do terceiro milênio. A pecuária tem um importante papel dentro do agronegócio neste período, com aproximadamente 8,5% do PIB total (Figura 2).

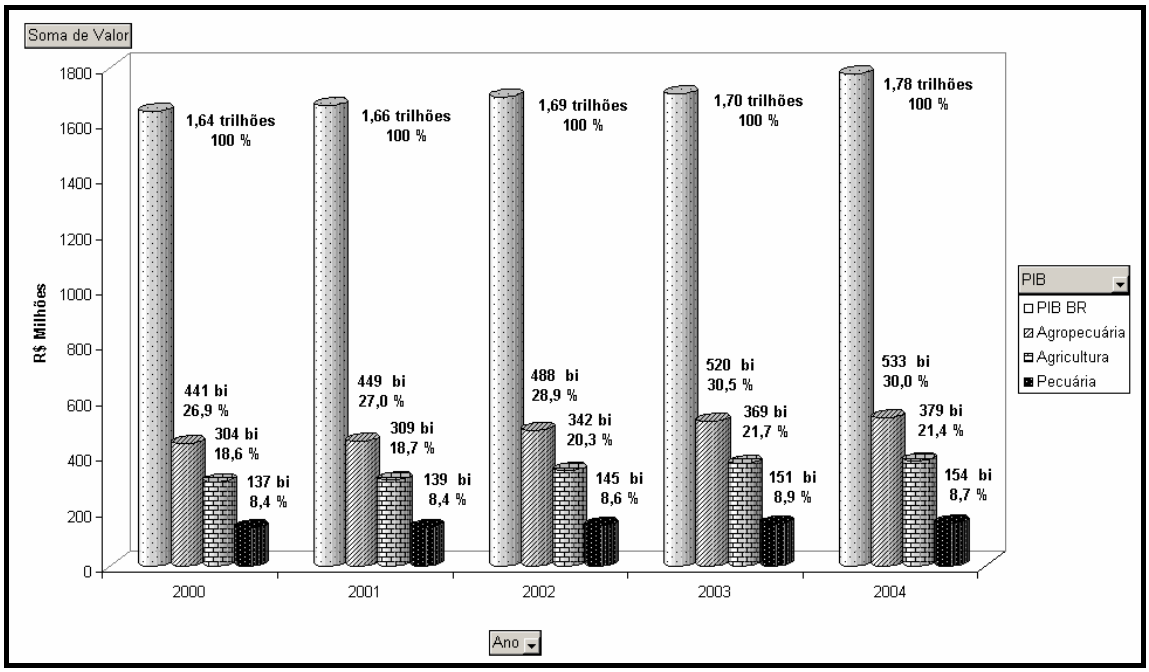


Figura 2 – PIB brasileiro e representatividade do agronegócio.
Fonte: Extraído de Centros de Estudos Avançados em Economia Aplicada [CEPEA] (2005).

A cadeia produtiva da bovinocultura de corte representa uma grande fonte de geração de divisas ao Brasil por meio das exportações. No ano de 2005, do total de US$ 118 bilhões gerados pelas exportações, o agronegócio foi responsável por US$ 46 bilhões (38%), especificamente a cadeia produtiva da bovinocultura de corte foi responsável por US$ 6 bilhões (5% do total e 13% do agronegócio) (Instituto de Economia Aplicada [IEA], 2006). Nestas estatísticas, estão apenas os produtos do setor alimentício e de vestuário, o impacto desta cadeia produtiva é, ainda, muito maior, já que ela fornece matéria-prima para 49 segmentos da indústria (Serviço de Informação da Carne [SIC], 2006a), dentre eles: construção civil, química, cosméticos, farmacêutica, automobilística e aeronáutica.

A pecuária de corte produz, como principal produto, um dos alimentos mais completos e saudáveis da alimentação humana, a carne bovina. A ingestão diária de apenas 100 g de carne bovina magra supre um homem adulto em 50% de suas necessidades de proteínas, além de outros nutrientes, com uma baixa taxa de gordura e colesterol (Tabela 1).

Dados da Associação Brasileiras dos Criadores de Zebu [ABCZ] (2006) demonstram a pujança da Raça Nelore na pecuária nacional, do efetivo do rebanho nacional de mais de 160 milhões de cabeças, estima-se que a Raça Nelore represente em torno de 80% da força produtiva da indústria da carne no Brasil, estando presente como raça pura ou como composição de cruzamentos. Considerando que foram

importados, segundo dados oficiais, 6.283 animais, pode-se constatar a incrível adaptabilidade do Nelore ao ambiente brasileiro, constituído, em sua maioria, por clima tropical e vegetação de cerrado. A interação do Nelore neste ambiente é responsável pela alta competitividade de nossa pecuária de corte, apresentando um dos menores custos de produção do mundo, sendo 60% mais baixo que o australiano, 50% menor que o americano e apenas um terço do custo britânico.

Tabela 1 – Importância da carne bovina na alimentação.

Componente	*Necessidades [a]*	*Oferecimento [b]*	*Unidades*	*Percentual [c]*
Energia	2500	186	cal	7%
Proteína	60	30	g	50%
Lipídeos	80	15	g	19%
Colesterol[d]	300	90	mg	30%
Zinco	9,5	3,5	mg	37%
Ferro	8	1,6	mg	20%
Riboflavina	1,3	0,26	mg	20%
Niacina	16	5,28	mg	33%
Vitamina B12	2,4	1,92	µg	80%

a) Homem adulto de 19 a 24 anos; b) Em 100 g de carne magra cozida; c) Oferecimento/Necessidades; d) Ingestão máxima diária recomendada;
Fonte: Serviço de Inspeção da Carne [SIC] (2006b).

Visto a importância da pecuária brasileira e a contribuição da Raça Nelore, pode-se predizer que a aplicação de tecnologias que aumentem a produção e produtividade de forma sustentável desta atividade econômica, em especial, desta raça, proporciona grande impacto sócio-econômico em nossa sociedade.

1.2. Princípios do melhoramento genético

A finalidade de um programa de melhoramento genético é a de utilizar a variabilidade genética da população para aumentar a produção e a produtividade dos animais. Numa raça bovina de corte, o objetivo final seria um aumento da taxa de desfrute do rebanho, com conseqüente aumento quantitativo e qualitativo de produção de carne e outros derivados.

Vários autores (OLIVEIRA, 2003; LÔBO, 1992; FOULLEY et al., 1990; VAN VLECK, 1987) destacam que a avaliação genética inicia-se dentro da fazenda, pela adoção de boas práticas de manejo. A correta identificação dos animais e formação dos lotes de manejo possibilita a coleta de dados não tendenciosos e a conectabilidade entre grupos contemporâneos.

Dado o modelo genético:

$$P = A + D + I + E + GE$$, em que:

P: Efeito fenotípico;

A: Efeito aditivo dos genes;

D: Efeito de dominância dos genes;

I: Efeito epistático dos genes;

E: Efeito ambiental;

GE: Interação genótipo-ambiente.

Todos os dados coletados na fazenda são fenotípicos (*P*), por exemplo, peso do animal ao completar um ano de idade. A avaliação genética consiste da aplicação de uma metodologia estatística capaz de isolar o efeito aditivo dos genes (*A*), demonstrando o valor genético do animal, ou seja, a contribuição da genética deste animal no peso e que pode ser transmitida aos seus filhos.

O uso da metodologia dos modelos mistos com propriedades *BLUP* (*Best Linear Unbiased Prediction* ou Melhor Preditor Linear não Viesado), sob modelo animal (PEREIRA, 2004; VAN VLECK, 1993), é atualmente a maneira mais difundida de avaliação genética para predição das DEPs dos animais. As DEPs estimam a metade do valor genético do animal, ou seja, o efeito aditivo médio (*A*) dos genes de um animal transmitidos à sua prole por meio de seus gametas.

O *BLUP* também permite o cálculo da acurácia, ou seja, a remoção da incerteza associada ao valor da DEP. Acurácia é um valor no intervalo *[0;1]*, sendo influenciada pelo número de progênies avaliadas que o animal deixou no rebanho e o número de rebanhos em que o animal deixou progênies avaliadas. Quanto maior forem estes valores, maior será o valor da acurácia (PEREIRA, 2004; ALBUQUERQUE et al., 2003; CARNEIRO et al., 2001).

Também é determinado o coeficiente de endogamia (F), definido como a probabilidade de serem idênticos dois alelos, no zigoto consangüíneo, devido ao parentesco dos pais (PEREIRA, 2004). Alto coeficiente de endogamia pode produzir perdas econômicas em virtude da perda de fertilidade, expressão de enfermidades autossômicas recessivas e perdas de alelos úteis nas populações.

A avaliação genética é uma “fotografia” do potencial genético de um rebanho, para que haja melhoramento genético, são necessários os corretos processos de seleção e acasalamento dos animais. Segundo Lush (1964), faz-se seleção artificial

quando se permite que alguns indivíduos da população produzam mais filhos do que outros, além de retirar os indesejáveis do processo reprodutivo (assim, aqueles que nascem, mas não têm oportunidade de reproduzir-se, não podem afetar a composição da população futura). Selecionando os melhores animais para serem reprodutores, a cada geração, a freqüência dos genes desejáveis vai se aproximando da meta desejada pelo selecionador do rebanho. Como conseqüência do correto processo de seleção, temos a média da característica selecionada da geração filial melhor que a geração paterna (Figura 3).

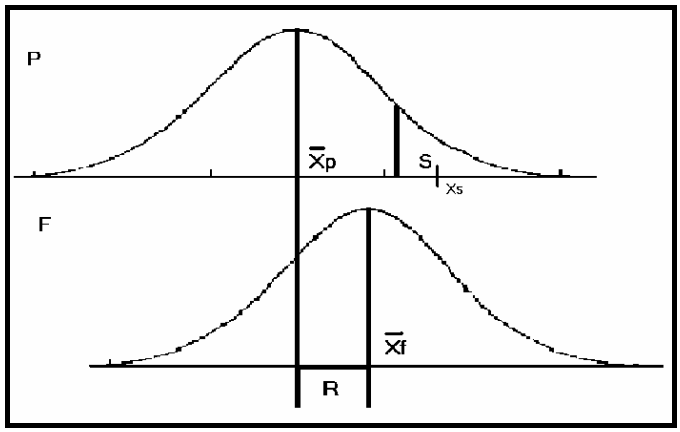


Figura 3 – Distribuição de uma característica hipotética em duas gerações, paterna (P) e filial (F), de um rebanho sob seleção, assinalando a média da geração paterna ($\bar{X}_p$), a média dos indivíduos selecionados ($\bar{X}_s$), a média da geração F ($\bar{X}_f$) e a resposta à seleção (R).
Fonte: Adaptado de Falconer e Mackay (1994, p. 199).

Embora um programa de melhoramento genético vise um equilíbrio entre as características de peso, fertilidade, reprodutivas, de habilidade maternal e de carcaça, as DEPs proporcionam ao criador a flexibilidade de criar um índice (conjunto de características) que ele considera útil ao seu rebanho. Os objetivos e critérios de seleção são pessoais e diferentes para cada criador, pois depende da expectativa do mercado consumidor (BERGMANN, 2003). Por exemplo, o criador que vende bezerros dará mais ênfase ao peso à desmama, já o criador que vende animais para o frigorífico, dará mais ênfase ao peso ao sobreano.

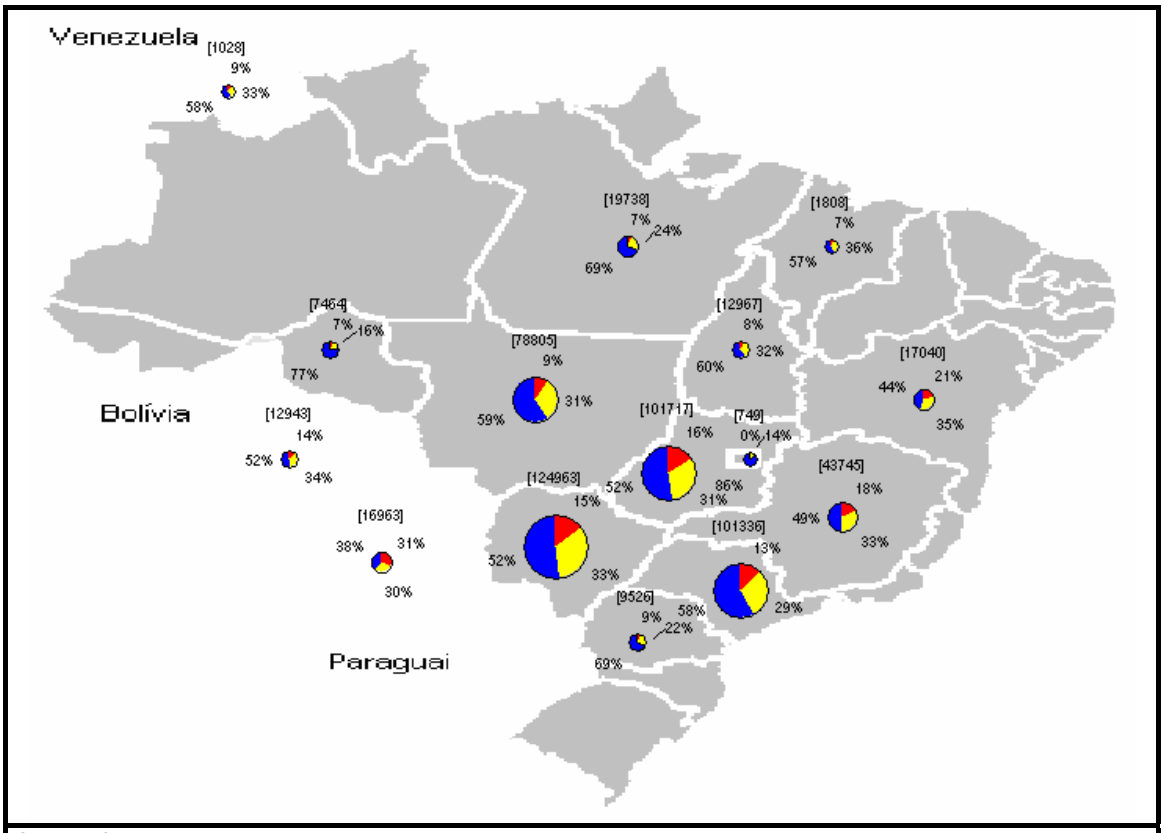
1.3. PMGRN – Nelore Brasil

O PMGRN – Nelore Brasil nasceu em 1988, quando dois criadores colocaram dados de seus animais à disposição do Grupo de Genética, Melhoramento Animal e Computação do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (GEMAC-DG-FMRP-USP). Em 1995, foi publicada a primeira Avaliação Genética de Touros, Matrizes e Animais Jovens (LÔBO et al., 2005). A partir de 1996 foi criada a Associação Nacional de Criadores e Pesquisadores (ANCP), com a finalidade de gerenciar o PMGRN – Nelore Brasil (Associação Nacional de Criadores e Pesquisadores [ANCP], 2006).

O PMGRN – Nelore Brasil publica, anualmente, o Sumário da Avaliação Genética de Touros e Matrizes e, semestralmente, disponibiliza o Sumário Eletrônico na *homepage* da ANCP (www.ancp.org.br). De modo geral, os sumários contêm os resultados das avaliações genéticas e disponibilizam aos criadores, o ordenamento dos animais em função de seus méritos genéticos para as diferentes características consideradas (TONHATI; MARCONDES; LÔBO, 2003), além de outras informações como a metodologia empregada, descrição das características avaliadas, critérios de inclusão dos animais e progresso genético do rebanho avaliado.

Uma das maiores contribuições da genética quantitativa é o cálculo do ganho genético obtido por estratégia de seleção. Com estas informações é possível orientar de maneira mais efetiva, o programa de melhoramento genético, predizer o sucesso do esquema seletivo adotado e decidir, com base científica, estratégias de seleção mais eficazes (CRUZ; REGAZZI, 1997). Na avaliação genética de 2005, o PMGRN – Nelore Brasil obteve ganho genético favorável para 9 das 11 características calculadas (MP120, DP120, DP365, DP450, DPE365, DPE450, DIPP, DPAC e MGT).

O PMGRN – Nelore Brasil, desde sua criação, experimentou um crescimento exponencial no número de animais avaliados e rebanhos participantes, além de um crescimento geográfico. Técnicas de visualização em mineração de dados (descritas na Seção 3.5) proporcionam, de maneira clara e objetiva, a idéia deste crescimento (Figura 4).



Legenda:
Cores por ano: ··· 1990-1994, ··· 1995-1999, ··· 2000-2004;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

Figura 4 – Distribuição dos animais avaliados pelo PMGRN – Nelore Brasil em função da área geográfica (estado brasileiro ou país da América Latina) e porcentagem de animais nascidos em função do período compreendido.

O grande número de animais avaliados, o rápido crescimento e a imensidão territorial ocupada pelo PMGRN – Nelore Brasil proporciona evolução da produção e produtividade da Raça Nelore, pela oferta de touros e matrizes de grande potencial genético. Porém, exige técnicas e ferramentas analíticas avançadas, para que seja possível administrar esse rebanho e continuar a proporcionar o ganho genético, que estão descritos na Seção 1.4.

Os ganhos genéticos do PMGRN – Nelore Brasil cumprem o objetivo de fornecer suporte ao criador para realizar a seleção e acasalamento dos animais, conforme descritos na Seção 1.2. Dada sua grandeza, ele, com certeza, causa grande impacto na produção e produtividade da cadeia da bovinocultura de corte descrita na Seção 1.1, trazendo benefícios sócio-econômicos.

1.4. Motivação e relevância do uso da tecnologia da informação pelo PMGRN – Nelore Brasil

O rápido crescimento da base de dados do PMGRN – Nelore Brasil, observado na Figura 4, segue uma tendência mundial, em que vários autores (EICK, 2000; GOEBEL; GRUENWALD, 1999) estimam que, a quantidade de informações disponíveis no mundo duplica, em média, a cada 20 meses.

O PMGRN – Nelore Brasil experimentou, na última década, os dois tipos de crescimento quantitativo dentro de um sistema de produção:

- **Crescimento horizontal:** Adesão de novos rebanhos avaliados;
- **Crescimento vertical:** Propagação de progênies dos animais já avaliados.

Os crescimentos horizontal e vertical demandam uma grande capacidade de armazenamento de dados, sendo necessários investimentos na capacidade de processar e utilizar os mesmos, para evitar o fenômeno batizado de “tumba de dados” (FAYYAD; UTHURUSAMI, 2002): *coleções de dados que são efetivamente armazenados para descansarem em paz e nunca mais serem acessados*.

Segundo Rezende (2003), vivemos na era do conhecimento, em que o nível dos indivíduos e da empresa é fator determinante de sua sobrevivência, bem como, o valor não está no domínio da informação, mas sim em como trabalhar o conhecimento relacionado a esta informação. Partindo desta premissa, o PMGRN – Nelore Brasil necessita de uma metodologia de análise da exploração, para transformar seus dados em vantagem competitiva aos criadores (Figura 5).

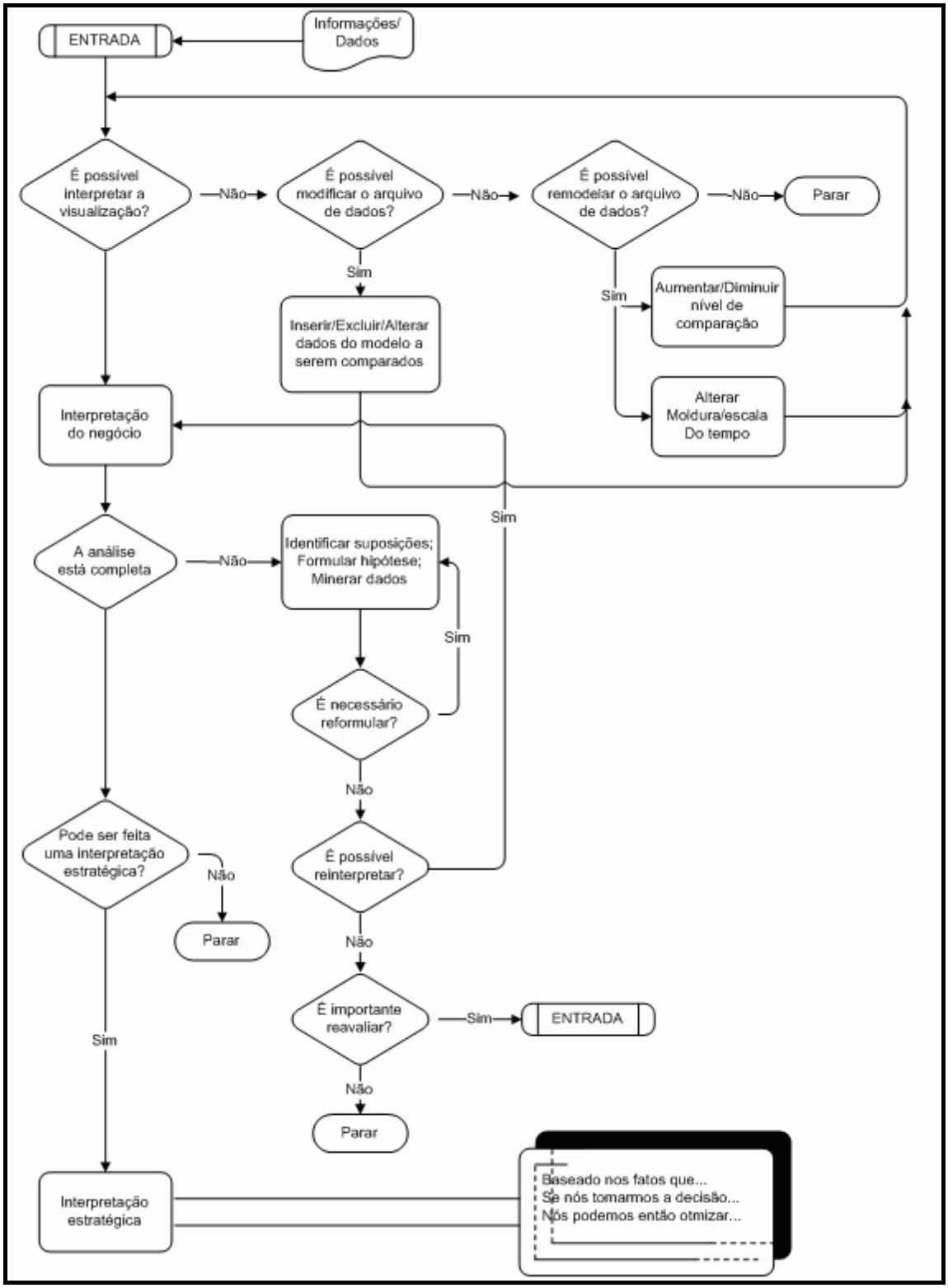


Figura 5 – Metodologia: dos dados à vantagem competitiva.

Fonte: Adaptado de Inmon, Terdeman e Imhoff (2000, p. 81).

O desenvolvimento de banco de dados, ferramentas analíticas e *links* de comunicação, além do treinamento da equipe de recursos humanos do PMGRN – Nelore Brasil podem conduzir a ganhos econômicos com utilização e venda dos

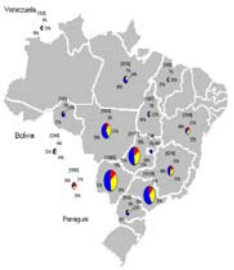
animais avaliados e geração de lucro. Para que haja este aumento de competitividade, as tecnologias devem ser amigáveis e estarem disponíveis a vários grupos de usuários como: administração e finanças, desenvolvimento e venda de produtos, prestação de serviços e assistência técnica, marketing e promoções (AMARAL, 2001). Dentro do programa de melhoramento genético existem todos estes grupos de usuários descritos.

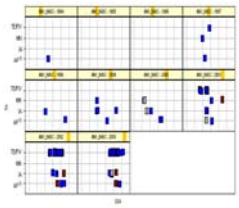
Lôbo (2005) descreve o histórico do uso da tecnologia da informação pelo PMGRN – Nelore Brasil. Em 1997, foi estabelecido o protocolo de parceria técnica entre o GEMAC-DG-FMRP-USP e o Laboratório de Inteligência Computacional do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (LABIC-ICMC-USP), para desenvolvimento de pesquisas na área de tecnologia da informação, buscando formar recursos humanos especializados em mineração de dados*, OLAP* e *Data Warehouse* (definições dos termos em Revisões Bibliográficas). Dentre as principais linhas de pesquisas deste grupo inter-unidades, cabe destacar:

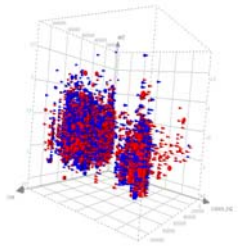
- Análise global da base de dados do PMGRN – Nelore Brasil utilizando, técnicas de visualização;
- Aplicação de métodos para identificação de falhas no manejo dos rebanhos;
- Implementação de um *Data Warehouse* para manuseio eficiente da grande bases de dados do PMGRN – Nelore Brasil;
- Desenvolvimento de ferramentas para disponibilizar conhecimento aos criadores e apoio à tomada de decisão.

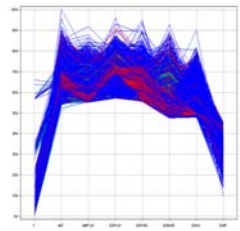
Como resultado desta parceria houve a implementação do *Sistema Nelore Business Intelligence*, um ambiente analítico que possibilita a extração de informações e conhecimentos da base de dados do PMGRN – Nelore Brasil (MARQUES, 2002).

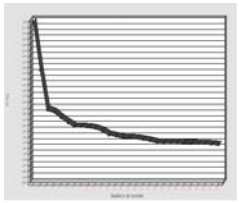
Agora chegou o momento de utilizar o *Nelore Business Intelligence* conforme a lógica de Knowles (KNOWLES, 1996), extraindo informações e conhecimentos em prol de vantagens competitivas ao PMGRN – Nelore Brasil.











OBJETIVOS

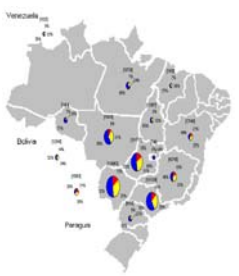
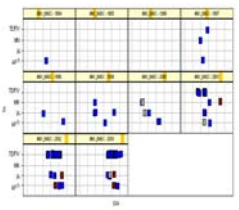
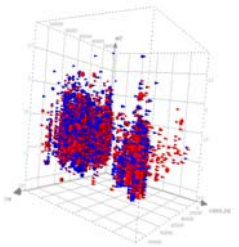
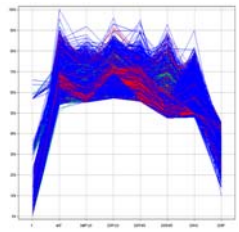
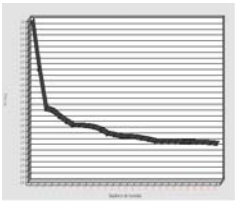
2. OBJETIVOS

2.1. Objetivos gerais

Extrair, da avaliação genética do PMGRN – Nelore Brasil, informações e conhecimentos com aplicativos de Processamento Analítico *On-Line* (*On-Line Analytical Processing – OLAP*) e de mineração visual de dados, respectivamente, utilizando um *Data Warehouse* (*Nelore Business Intelligence*) como fonte única de dados.

2.2. Objetivos específicos

- Caracterizar a estrutura populacional do rebanho avaliado pelo PMGRN – Nelore Brasil;
- Identificar as vias de fluxo gênico na Raça Nelore;
- Extrair padrões de seleção e acasalamento das fazendas participantes.

REVISÃO BIBLIOGRÁFICA

3. REVISÃO BIBLIOGRÁFICA

3.1. Princípios da tecnologia da informação

Esta Seção apresenta conceitos básicos da tecnologia da informação e exemplifica como estes conceitos, aplicados ao PMGRN – Nelore Brasil, são capazes de gerar uma nova perspectiva administrativa, fundamental ao sucesso da organização.

Dados, **informação** e **conhecimento**, em tese esses três termos parecem representar a mesma coisa, porém eles possuem conceitos diferentes e podem determinar o sucesso ou fracasso de uma empresa ou organização.

Num experimento pioneiro com mineração de dados, o Cassino Harrah's, em Las Vegas, minerou seu banco de dados (com o perfil de 16 milhões de clientes) e extraiu um conhecimento valiosíssimo: os apostadores que gastam entre 100 e 500 dólares, numa visita ao cassino, correspondem a apenas 30% de toda a clientela, mas contribuem com 80% das receitas. Com estratégias agressivas de marketing para atrair esse filão mais rentável (são oferecidos almoços, shows e apostas grátis), o cassino afirma ter dobrado seu faturamento em um ano (GUIZZO, 2001).

Para compreender a importância de utilizar a tecnologia da informação é preciso entender a hierarquia do conhecimento humano (GOLDSCHMIDT; PASSOS, 2005; REZENDE, 2003):

- **Dados:** São elementos puros, quantificáveis sobre um determinado evento. São valores ou descrições que uma variável assume dentro de um banco de dados. Os dados não fornecem embasamento para o entendimento de uma situação, porém são os componentes básicos a partir do qual uma informação será gerada;
- **Informação:** É o dado inserido num contexto (situação que está sendo analisada). É preciso estabelecer um parâmetro de comparação, ou seja, criar a informação;
- **Conhecimento:** É a descoberta das causas que influenciam variações nos resultados. O conhecimento refere-se à habilidade de criar um modelo mental que descreva o objeto e indique as ações a implementar, as decisões a tomar. O processo de gerar conhecimento resulta de um processo no qual uma informação é comparada à outra e combinada em muitas ligações

(hiperconexões) úteis com significado. Isso implica que o conhecimento é dependente de nossos valores e nossa experiência e sujeito às leis universalmente aceitas. O conhecimento proporciona compreensão, análise e síntese, necessários para a tomada de decisões inteligentes.

Ilustrando melhor estes conceitos, temos uma situação hipotética, em que um criador tem como critério de seleção, utilizar touros com alto potencial genético para produzir bezerros pesados à desmama. Esse criador se depara com três opções de sêmen à venda (Tabela 2), qual o melhor touro?

Tabela 2 – Situação hipotética de três touros para comparação de DEPs.

Touro	*DP120 (Kg)*	*Acurácia*	*Percentil*	*NF120*
A	9,00	20%	0,1%	1
B	8,00	48%	0,1%	40
C	2,00	52%	25,0%	45

Comparando os **dados** de DEP para efeito direto no peso aos 120 dias (DP120) das três opções de touros da Tabela 2 e baseando no objetivo de seleção do criador, o touro *A* é a melhor opção.

Com o objetivo de minimizar riscos de investimento, a fazenda tem a meta de utilizar touros com acurácia igual ou maior a 50% (contexto). Este parâmetro de comparação permite a escolha baseada na **informação**. Neste caso, o touro *C* é a melhor opção.

Se o criador, por experiência, sabe que o número de filhos avaliados que um touro deixa no rebanho (NF120) interfere diretamente na acurácia e a diferença entre 40 e 45 filhos avaliados é mínima, então a comparação passa a ser baseada no **conhecimento** e como uma acurácia de 48% é próxima à meta da fazenda, ele faz a opção de comprar sêmen do touro *B*. O conhecimento proporcionou uma tomada de decisão inteligente, ou seja, o uso de um touro com alta DP120 (semelhante ao Touro *A*), com alta acurácia (semelhante ao Touro *C*). O Touro *B* conseguirá maximizar o ganho genético do rebanho com baixo risco de investimento.

Para a coleta de dados provenientes das fazendas, o PMGRN – Nelore Brasil utiliza o *Sistema Nelore* (*SisNe*). Este é um banco de dados operacional, convencional ou de produção, pois possui processamento de transações (*On-Line Transactional Processing – OLTP*).

Com o objetivo de realizar análises e aumentar a competitividade do PMGRN – Nelore Brasil, foi implementado o *Sistema Nelore Business Intelligence* (MARQUES,

2002). Ele é um sistema de inteligência empresarial (*Business Intelligence System – BIS*), proporcionando aos usuários acesso aos dados com grande capacidade analítica, otimizando suas operações (Tabela 3).

Tabela 3 – Diferenças entre sistema operacional e sistema de inteligência empresarial (*BIS*) quanto às expectativas do usuário.

Sistema operacional – SisNe	***BIS – Nelore Business Intelligence***
Consultas carregam pequeno volume de informação. Ex.: Genealogia de um animal.	Consultas trabalham com grandes blocos de informação. Ex.: Média das DEPs dos animais por estado para as safras 1994 e 2003.
Dados atualizados freqüentemente. Ex.: A cada pesagem dos animais.	Dados atualizados periodicamente. Ex.: A cada avaliação genética.
Entrada rápida de dados. Ex.: Quando uma fazenda envia dados da pesagem de seus animais, estes são rapidamente inseridos no sistema.	Não há entrada de dados, o sistema suporta apenas atualização e consulta.
Respostas imediatas. Ex.: Quando o usuário precisa consultar o ano de nascimento de um animal, ela demora frações de segundos.	Respostas rápidas. Ex.: Quando o usuário precisa consultar uma tendência genética, a consulta demora de 1 a 10 minutos.
Padrão de utilização do sistema pelo usuário é relativamente previsível. Ex.: O usuário sempre quer consultar dados de um determinado animal.	Usuário necessita de liberdade de uso do sistema, para realizar análises complexas. Ex.: O usuário necessita, a cada consulta, de diferentes estatísticas.
As consultas retornam dados e relações entre dados. Ex.: Listagem de todas as progênies de um determinado touro.	As consultas retornam conceitos, informações e regras. Ex.: Média do peso dos animais a desmama em função do mês do nascimento.

Fonte: Adaptado de Rezende(2003); Corey et al. (2001).

O *Nelore Business Intelligence* é atualizado pelo *SisNe* após a avaliação genética. A atualização deste *BIS*, pelo sistema operacional, obedece ao ciclo de interação entre diferentes bancos de dados descrito por Inmon, Terdeman e Imhoff (2001). Neste ciclo, dados, que nada significam, são agrupados de maneira a formar elementos capazes de fornecer significado, quando sujeitos a exploração ou mineração, produzindo informações e conhecimentos (COREY et al., 2001).

3.2. *Data Warehouse*

Esta Seção explana sobre o *Data Warehouse* e o *Nelore Business Intelligence*.

O desenvolvimento da tecnologia da informação tem permitido às empresas manusearem grandes volumes de dados e atingirem um alto índice de globalização, com o uso de redes, viabilizando operações em nível mundial. A todo instante, dados sobre os mais variados aspectos dos negócios da empresa são gerados e armazenados, e passam a fazer parte dos recursos de informação da empresa. Em

princípio podemos encarar isso como um ponto a favor da empresa, mas que na verdade pode constituir um problema quando esses dados encontram-se espalhados em diversos sistemas, e exigem um esforço grande na tentativa de integrá-los para que possam ter alguma utilidade (COME, 2001). O primeiro passo para explorar esses dados, que geralmente encontram-se espalhados por diversos sistemas, é integrá-los num ambiente organizacional.

O *Data Warehousing* é justamente o processo da área de tecnologia da informação, que objetiva satisfazer as necessidades dos usuários quanto ao armazenamento e acesso aos dados, que geram visões multidimensionais e dão amplo apoio ao processo de tomada de decisão. Este é o processo de construção, acesso e manutenção do *Data Warehouse*. Segundo Rezende et al. (2003), *Data Warehousing* é um processo e *Data Warehouse* é seu produto.

O *Data Warehouse* é um banco de dados (um grande repositório de dados) com função analítica, ou seja, sua função é proporcionar aos usuários, uma única fonte de informação a respeito dos seus negócios, servindo também, como ferramenta de apoio ao processo de extração de informação. Inmon (1997) descreve o *Data Warehouse* como um conjunto de dados orientado por assunto, integrado, não volátil, variante no tempo, no apoio de decisões gerenciais. O *Nelore Business Intelligence* obedece a essas quatro característica do *Data Warehouse* (Tabela 4).

Tabela 4 – Relação entre as características do *Data Warehouse* e exemplo no *Nelore Business Intelligence*.

Característica	***Data Warehouse***	***Nelore Business Intelligence***
Orientado por assunto	Orientado ao redor do principal assunto da organização.	Organizado ao redor das avaliações genéticas dos animais.
Integrado	Convenção consistente de nomes, forma consistente das variáveis, estrutura consistente de códigos, atributos físicos e dos dados.	Pesos sempre expressos em quilogramas (Kg) e perímetro escrotal, sempre em centímetros (cm).
Variante no tempo	Todo dado é exato em algum momento do tempo, os valores são históricos.	O peso da vaca ao parto (PVP) de uma matriz aferido a cada parição é armazenado segundo a sua data.
Não volátil	Não existem operações de alteração e exclusão de dados. Somente duas espécies de operações ocorrem no *Data Warehouse* - a carga inicial do dado, e o acesso ao dado.	Relaciona o novo peso do animal a uma nova data de pesagem, ao invés de ser substituído.

O *Nelore Business Intelligence* foi desenvolvido utilizando o Sistema Gerenciador de Banco de Dados (SGBD) *Oracle 8i* (Oracle, 1999) e a Ferramenta de Engenharia de Software (*Computer-Aided Software Engineering – CASE*) e de

Extração, Transformação e Povoamento (*Extraction, Transformation and Load – ETL*), *Oracle Warehouse Builder* (Oracle, 1999).

Data Marts são estruturas de consulta de alto desempenho, em que os usuários finais devem fazer suas consultas. São conceitos lógicos e não físicos, ou seja, não são mais um repositório de dados e sim, um conjunto de estruturas, às quais contêm dados em formatos que tornam mais fácil e rápido o acesso (COREY et al., 2001).

A arquitetura do *Nelore Business Intelligence* permite a relação entre sistema operacional, *Data Warehouse*, *Data Mart*, consulta *OLAP* e mineração de dados (Figura 6).

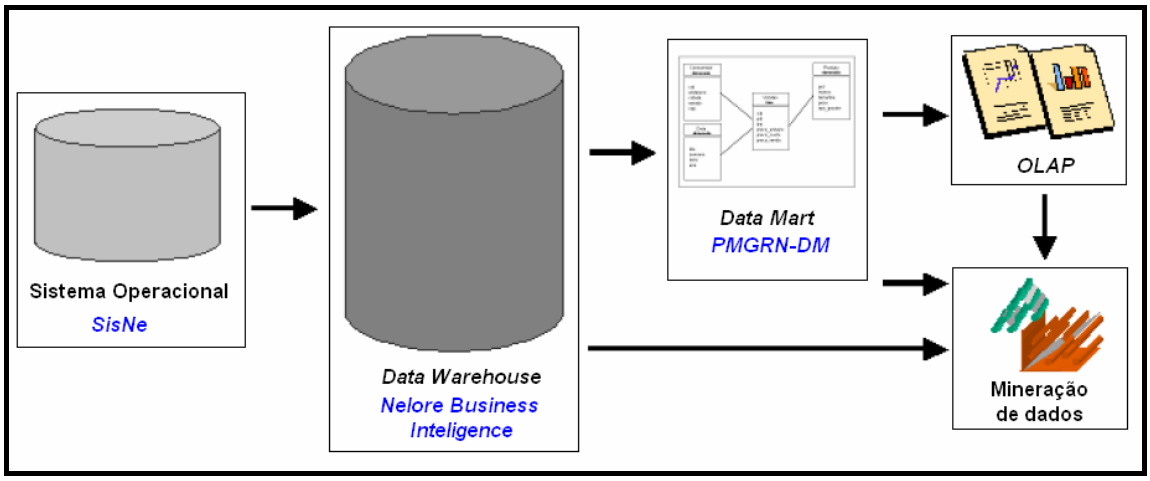


Figura 6 – Arquitetura do *Nelore Business Intelligence*.
Fonte: Adaptado de Marques (2002, p. 69).

Na Figura 6, podemos ver que se trata de uma topologia *Data Marts* dependentes, em que estes são criados a partir de um *Data Warehouse*. Até o momento, foi implementado no *Nelore Business Intelligence*, apenas o *Data Mart* referente à avaliação genética, chamado de *Data Mart* do Programa de Melhoramento Genético da Raça Nelore (*PMGRN-DM*). Essa arquitetura permite implementar novos *Data Marts* para outros departamentos da organização (COREY et al., 2001; INMON; TERDEMAN; IMHOFF, 2001; GATZIU; VAVAOURAS, 1999; GARDNER, 1988), como: financeiro, marketing, consultorias às fazendas, entre outros.

O *PMGRN-DM* utiliza a tecnologia de Banco de Dados Relacionais (*RDBMS*[1]), com modelagem multidimensional e esquema constelação composto de quatro tabelas de fatos e seis tabelas de dimensões (Figura 7). As dimensões possuem diferentes atributos e hierarquias (Tabela 5). A tecnologia *RDBMS* proporciona Processamento Analítico *On-Line* do tipo Relacional (*ROLAP*), que será visto com mais detalhes na Seção 3.3.

O *SisNe* é altamente normalizado[2] e o *Nelore Business Intelligence*, desnormalizado[3] em esquemas estrelas. Enquanto que a modelagem tradicional dos sistemas operacionais são altamente normalizadas para assegurarem o cumprimento de restrições e evitarem as redundâncias de informações, os sistemas analíticos possuem a modelagem multidimensional desnormalizadas, que aceleram o desempenho das consultas (COREY et al., 2001; GOLDSCHMIDT; PASSOS, 2005).

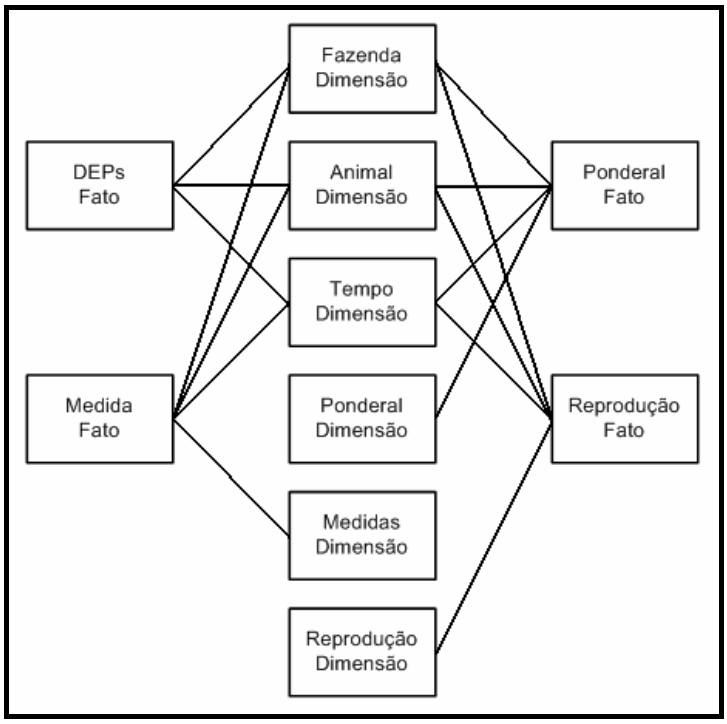


Figura 7 – Esquema constelação do *PMGRN-DM*.
Fonte: Marques (2002, p. 75)**.**

[1] ***RDBMS*:** A estrutura de banco relacional geralmente consiste de um esquema Estrela (*Star*) ou algum variante.

[2] **Normalizado:** Uma estrutura de banco de dados normalizada permite o armazenamento de dados de forma flexível, possuindo várias tabelas e chaves primárias.

[3] **Desnormalizado:** Uma estrutura de banco de dados desnormalizada proporciona consultas de alto desempenho, possuindo poucas tabelas e chaves primárias.

Tabela 5 – Dimensões, atributos e hierarquias do *PMGRN-DM.*

Dimensão	*Atributos*	*Hierarquia*
Fazenda	*Id_Fazenda*[1], *Código, Nome Fazenda, Nome Criador, Município, Estado, Pessoa, Pesa ao Nascer, Dt_Inserção*[3], *Dt_Atualização*[4], *Flag*[5]	Estado → Município → Criador → Fazenda
Animal	*Id_Animal*[1], *CGA*[2], *Nome, Sexo, Raça, Categoria, Sit_Nasc, Situação, Oco_Parto, Dt_Inserção*[3], *Dt_Atualização*[4], *Flag*[5]	
Tempo		Ano → Mês → Dia
Ponderal	*Id_ponderal, M120, M240, M365, M455, M550, M730, MPDB, MPVD, Dt_inserção*[3]	
Medidas	*Id_Animal*[1], *CGA*[2], *Dt_Peso*[2], *Manejo, Situação, Dt_Inserçao*[3]	Manejo → Situação
Reprodução	*Nar*[2], *Tipo Aças, Manejo, Dt_Inserção*[3]	Tipo de acasalamento → Manejo da vaca ao parto

(1) Chaves substitutas; (2) Chaves originais do *SisNe*; (3) Data em que o registro foi inserido; (4) Data em que o registro foi atualizado; (5) Indica se o registro está ativo ou não;
Fonte: Adaptado de Marques (2002).

O esquema constelação, representado pela Figura 7, integra quatro esquemas estrelas, um para cada relacionamento de uma tabela fato com suas tabelas dimensões. O esquema estrela é a tentativa de se chegar o mais próximo de um arquivo simples, com dados ligeiramente normalizados, que asseguram enormes ganhos de desempenho durante as consultas *OLAP* (COREY et al., 2001; Oracle, 1999).

No esquema estrela, a tabela de fatos contém valores que são mensuráveis (Ex.: Peso do animal) e as dimensões contém informações descritivas a respeito destes valores (Ex.: Estado), em resumo, as dimensões dão significados aos fatos (COREY et al., 2001; Oracle, 1999; GOLFARELLI; MAIO; RIZZI, 1998). Durante uma consulta, fatos e dimensões funcionam de modo análogos a uma fração, em que o numerador são os fatos e o denominador, as dimensões. Ex.: Se desejássemos consultar a média da DP120 por estado e safra, teríamos a consulta segundo a equação abaixo:

$$Consulta = \frac{Fatos}{Dimensões} = \frac{AVG(DDPP120)}{(Estado) \times (Ano\_do\_Nascimento)}$$

As hierarquias permitem ao usuário realizar as funções manobra a cima (*drill-up*) e manobra abaixo (*drill-down*), descritas na Seção 3.3, em que é possível viajar

por diferentes níveis de granularidade[4] dos dados, por exemplo, a mesma consulta da equação acima poderia ser realizada por *Município*.

Com todas essas configurações descritas, o *Nelore Business Intelligence* permite analisar quatro grupos de objetos:

- ***DEP*:** Resultados da avaliação genética dos animais, como DEPs, acurácias, Mérito Genético Total (MGT), Coeficiente de Endogamia (F), Número de Filhos (NF) e Número de Rebanhos (NR);
- ***Medidas*:** Pesagens e aferições de perímetro escrotal (PE);
- ***Ponderal*:** Padronização dos pesos e PE a idades pré-determinadas;
- ***Reproducao*:** Informações reprodutivas dos animais.

3.3. Consulta *OLAP*

Esta Seção descreve as funcionalidades do aplicativo *OLAP* utilizado, o *Oracle Discoverer 4* (Discoverer, 2000b).

O processamento *OLAP* é constituído de um conjunto de tecnologias capazes de resumir e analisar grandes volumes de dados. Os sistemas *OLAP* permitem aos usuários, verem medidas do desempenho organizacional decompostas pelas dimensões dessas medidas. O aplicativo *OLAP* faz a interface do usuário com o *Data Warehouse*, capacitando-o a realizar as análises multidimensionais, ou seja, ele é o aplicativo de inteligência empresarial (COREY et al., 2001; COLLIATE, 1996).

O aplicativo *OLAP* proporciona o processo de exploração, definido por Inmon, Terdeman e Imhoff (2001), como a atividade de procurar informações que concederão significativa vantagem de negócio a partir de dados que são reunidos pela corporação em seu *Data Warehouse*.

O *Oracle Discoverer* é uma ferramenta do tipo *ROLAP,* pois acessa um *RDBMS*. A estratégia *ROLAP* oferece vários seguintes benefícios (COREY et al., 2001; COLLIATE, 1996) que foram utilizadas no *Nelore Business Intelligence*:

- **Suporte para um crescimento praticamente ilimitado:** O número de animais avaliados pelo PMGRN – Nelore Brasil cresce rapidamente;

[4] **Granularidade** ou **granulosidade:** Nível de agregação dos dados. Ex.: DEPs por animal, fazenda, município, estado ou país.

- **Facilita o desempenho de carregamento:** Caso venha a ser implementado outros *Data Marts*, que exigem carregamento mais frenético, como por exemplo, controle financeiro com atualização mensal, sua agilidade elimina inconveniências de longo tempo de indisponibilidade do sistema, aos usuários;
- **Permite ampla liberdade de navegação pelas dimensões e hierarquias:** Proporciona ao usuário, analisar seus dados em diversos níveis de granularidade;

Segundo Corey et al. (2001), o recurso de Tabela Dinâmica do *Excel* (Excel, 2003) embora não suporte facilmente a exploração, aproxima-se de um ambiente *OLAP*.

3.3.1. Facilidade de uso

O *Oracle Discoverer* possui dois módulos distintos:

- **Edição de Administração (*Administration Edition*):** Voltado a administradores de sistema. É utilizado para construir áreas empresariais, definir o perfil e acesso de usuários, personalizar agrupamentos e hierarquias dos dados, segundo perfis de usuários (Discoverer, 2000a);
- **Edição de Usuários (*User Edition*):** Voltado a usuários do sistema. Permite definir, criar e personalizar relatórios e gráficos de forma livre e flexível, com os dados definidos nas áreas empresariais (Discoverer, 2000b).

A Edição de Administração esconde do usuário final, toda a complexidade do *Data Warehouse* e seus *Data Marts*. O usuário final precisa apenas da Edição de Usuários para acessar ao *Data Warehouse* (*Nelore Business Intelligence*), portanto ele não precisa ter conhecimentos profundos de informática e sim, do domínio, que neste caso é o melhoramento genético.

3.3.2. Segurança

O *Data Warehouse* beneficia-se da segurança proporcionada pelo *Oracle 8i* e pelo *Oracle Discoverer Administration Edition*, fornecendo segurança no manuseio de banco de dados, lembrando que este é um patrimônio da organização ou instituição.

O administrador do sistema conta com o recurso de segurança baseado no papel, em que ele gerencia usuários[5], papéis[6], alistamentos[7] e perfis[8], distribuindo-os aos usuários com base nas responsabilidades de cada um (COREY et al., 2001).

Com a segurança baseada no papel, poderia ser oferecido a cada usuário do PMGRN – Nelore Brasil, uma senha de acesso ao *Nelore Business Intelligence*, assim ele poderia analisar apenas dados referentes aos animais das fazendas sobre sua responsabilidade. Este usuário, ao estabelecer conexão ao banco de dados, é interrogado pelo *Oracle Discoverer* em Usuário e Senha específicos (Figura 8).

Figura 8 – Menu de conexão do *Discoverer* ao *Data Warehouse.*

3.3.3. Quatro tipos diferentes de relatórios

O Oracle Discoverer fornece ao usuário, quatro diferentes tipos de relatórios, ou seja, o modo de exibição do resultado da consulta (Figura 9):

- **Tabela:** Exibe os dados em linhas e colunas, essa é a representação de dados tradicional;
- **Tabela detalhada por página:** Permite a exibição de informações agrupadas pelos critérios especificados no item de páginas;

[5] **Usuários:** Pessoas que recebem uma senha para acessar o banco de dados;

[6] **Papéis:** Agrupamentos lógicos de um ou mais usuários, aos quais são concedidos privilégios de análises;

[7] **Alistamentos:** Processos de dar participações como membros em um papel para um ou mais usuários;

[8] **Perfis:** Conjuntos de limites de recursos que podem ser oferecidos aos usuários.

- **Tabela de referência cruzada:** Exibe dados multidimensionais e permite a criação de pivôs com as dimensões entre as linhas e colunas da tabela;
- **Tabela de referência cruzada detalhada:** Permite exibir informações agrupadas para critérios especificados no item de páginas, essa é a representação que permite aplicar todas as funções *OLAP*.

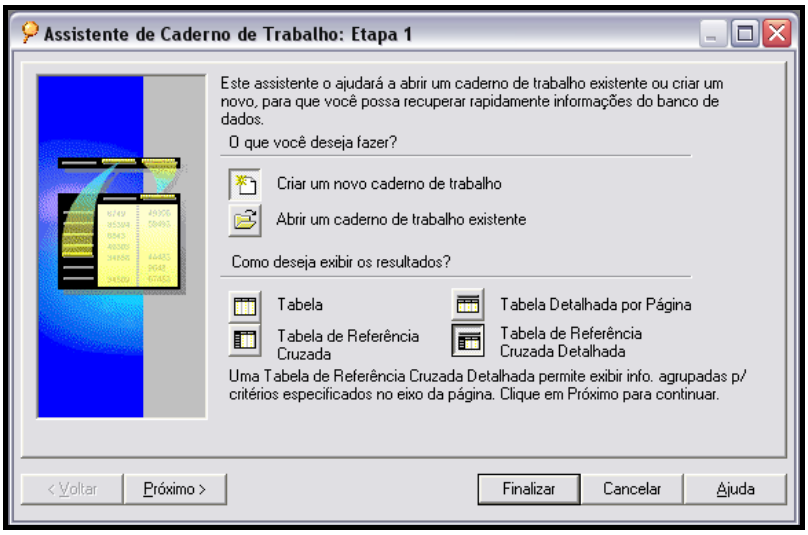


Figura 9 – Menu com tipos de relatórios.

3.3.4. Liberdade de seleção da área empresarial e objetos internos

O *Nelore Business Intelligence* possui uma única área empresarial, o *PMGRN-DM*, com 68 objetos para o grupo *DEP*, 28 para *Medidas*, 44 para *Ponderal* e 36 para *Reprodução*.

Se selecionássemos todos os objetos do menor grupo, *Medidas*, ele geraria no mínimo, 28 dimensões referentes aos objetos (isso se, para cada objeto da tabela fato, tivesse apenas uma operação, como a média aritmética). Um número tão grande de dimensões demandaria horas de processamento do computador para gerar o relatório.

Como usuários necessitam de agilidade no processamento de relatórios, a ferramenta *Oracle Discoverer* fornece total liberdade de seleção da área empresarial e dos objetos inerentes à análise (Figura 10). Com a escolha das dimensões corretas (inerentes à análise), a geração de um relatório geralmente leva menos que 10 minutos de processamento.

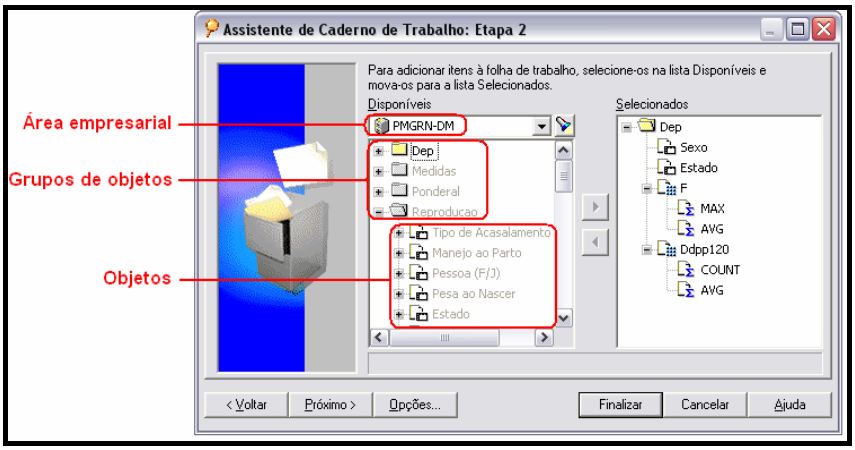


Figura 10 – Menu de seleção da área de trabalho, grupo de objetos e objetos.

3.3.5. Aplicação de filtros

A aplicação de filtros é outra solução ao problema de tempo de processamento, além de permitir ao usuário, a seleção, geográfica e temporal, de uma amostra desejada na análise. Para qualquer objeto selecionado ou não, é possível aplicar filtros (Figura 11). O usuário pode aplicar filtros durante a seleção dos dados ou após a geração de um relatório inicial.

No exemplo da Figura 11, foram selecionados apenas animais dos estados de Minas Gerais, Mato Grosso do Sul e São Paulo, além de nascidos nas safras de 2001 a 2003.

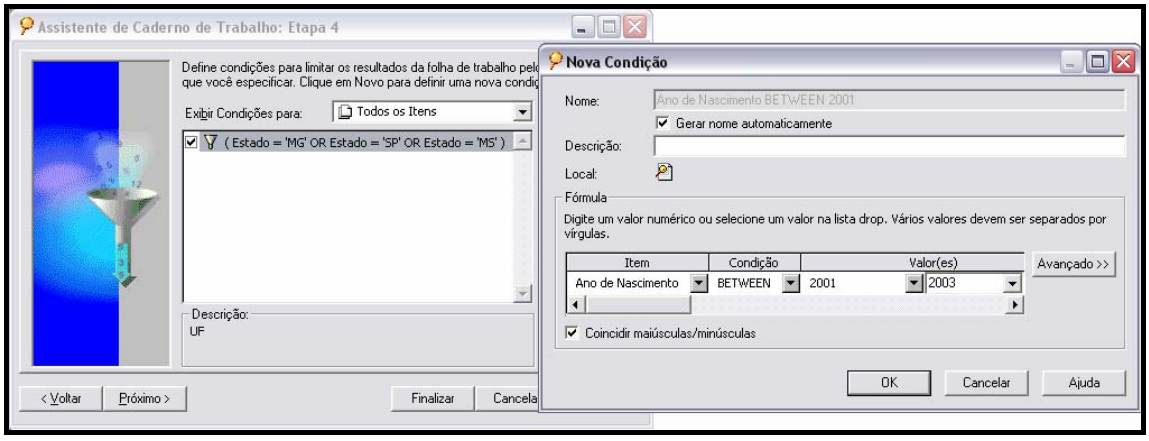


Figura 11 – Menu para aplicação de filtros.

3.3.6. Cálculos

O *Oracle Discoverer* oferece ao usuário, funções do tipo (Figura 12):

- **Analítico:** Funções estatísticas e agregadas para uma partição de dados. Ex.: *AVG* – retorna a média aritmética para a partição de dados selecionados;
- **Conversão:** Para converter os dados de um tipo para outro. Ex.: *HEXTORAW* – converte caracteres que contém dígitos hexadecimais em um valor bruto;
- **Data:** Para manusear datas. Ex: *MOUNTH_BETWEEN* – retorna o número de meses entre duas datas;
- **Grupo:** Funções estatísticas e agregadas para todo o objeto. Ex.: *CORR* – retorna o coeficiente de correlação de Pearson entre duas variáveis;
- **Numérico:** Para itens numéricos, transcedentais e pontos de flutuação. Ex.: *LN* – retorna o logaritmo natural de uma variável;
- **Outras:** Funções lógicas, entre outras. Ex.: *CASE* – função lógica de estrutura condicional do tipo *Se... Então... Senão...*;
- **String:** Operações com texto. Ex.: *CONCAT* – concatena duas variáveis.

O usuário pode incluir cálculos durante a seleção dos dados ou após geração de algum relatório.

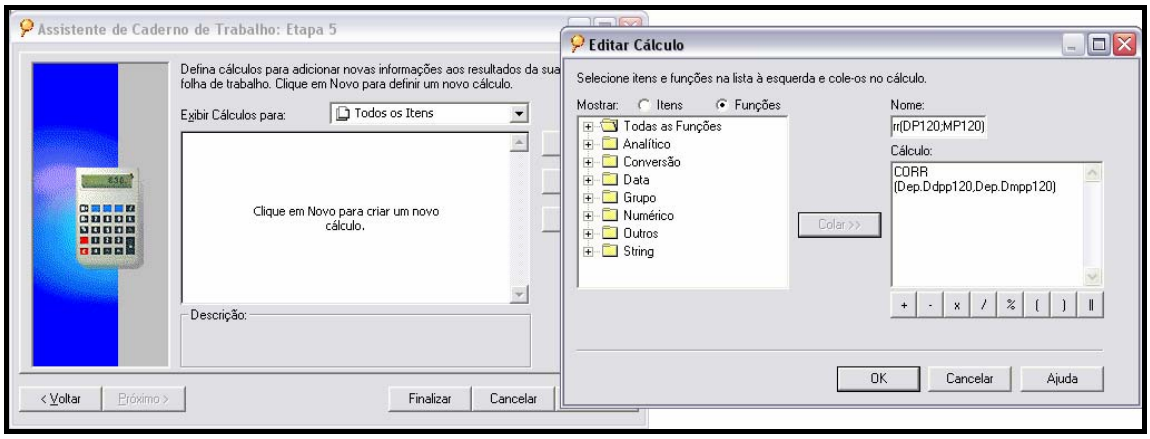


Figura 12 – Menu para editar cálculos.

3.3.7. Funções *OLAP*

As ferramentas apresentadas pelo *Oracle Discoverer* como cálculos e filtros, apesar de serem excelentes recursos para analisar dados, são semelhantes aos aplicativos de análises estatísticas clássicos. A característica que diferencia os aplicativos *OLAP* dos demais é sua capacidade em deixar o usuário livre para rotacionar os dados, trocar as dimensões de posição, extrair novas informações de um relatório previamente montado e definir vários modos de visualização (COREY et al., 2001).

A capacidade de trabalhar com multidimensionalidade é determinada pelas funções *OLAP* (Figura 13). A partir de uma consulta em tabela de referência cruzada detalhada (Figura 13 – Consulta original), com média da DP120, para alguns estados (MG, SP e MS) e algumas safras (2001 a 2003), foram testadas as seguintes funções *OLAP* (SHOSHANI, 1997):

- **Pivoteamento (*Pivoting*):** Função que muda a orientação dimensional de uma pesquisa. Pode ocorrer dentro de uma dimensão, na troca de dimensões ou troca de uma dimensão por um dos atributos do item de páginas. Ex.: Troca de posição das dimensões *Estado* e *Ano do Nascimento* (Figura 13 – Pivoteamento);
- **Rolamento (*Roll-up*):** Função que consiste da troca de hierarquias, fórmula ou operações dentro de uma dimensão. Ex.: Troca da fórmula média (*AVG*) por contar número (*COUNT*) na dimensão *DDPP120* (Figura 13 – Rolamento);
- **Repartimento (*Slicing*):** Função que fixa um valor simples em lugar de um ou mais atributos das dimensões. Ex.: Fixar o valor da dimensão *Sexo* em "Macho" (Figura 13 – Repartimento);
- **Manobra abaixo/acima (*Drill-down/up*):** Quando são instituídas hierarquias dentro de uma dimensão, esta função permite explorar diferentes níveis de granularidade dessa dimensão. Ex.: *Drill-down*, para aumentar o nível de detalhamento geográfico para o estado de MG, visualizando a média da dimensão *DDPP120* por Município (Figura 13 – Manobra abaixo/acima);
- **Manobra de junção (*Drill-across*):** Função que permite unir duas ou mais dimensões com o mesmo nível de detalhamento. Ex.: União da dimensão *Sexo* à dimensão *Ano do Nascimento* (Figura 13 – Manobra de junção);

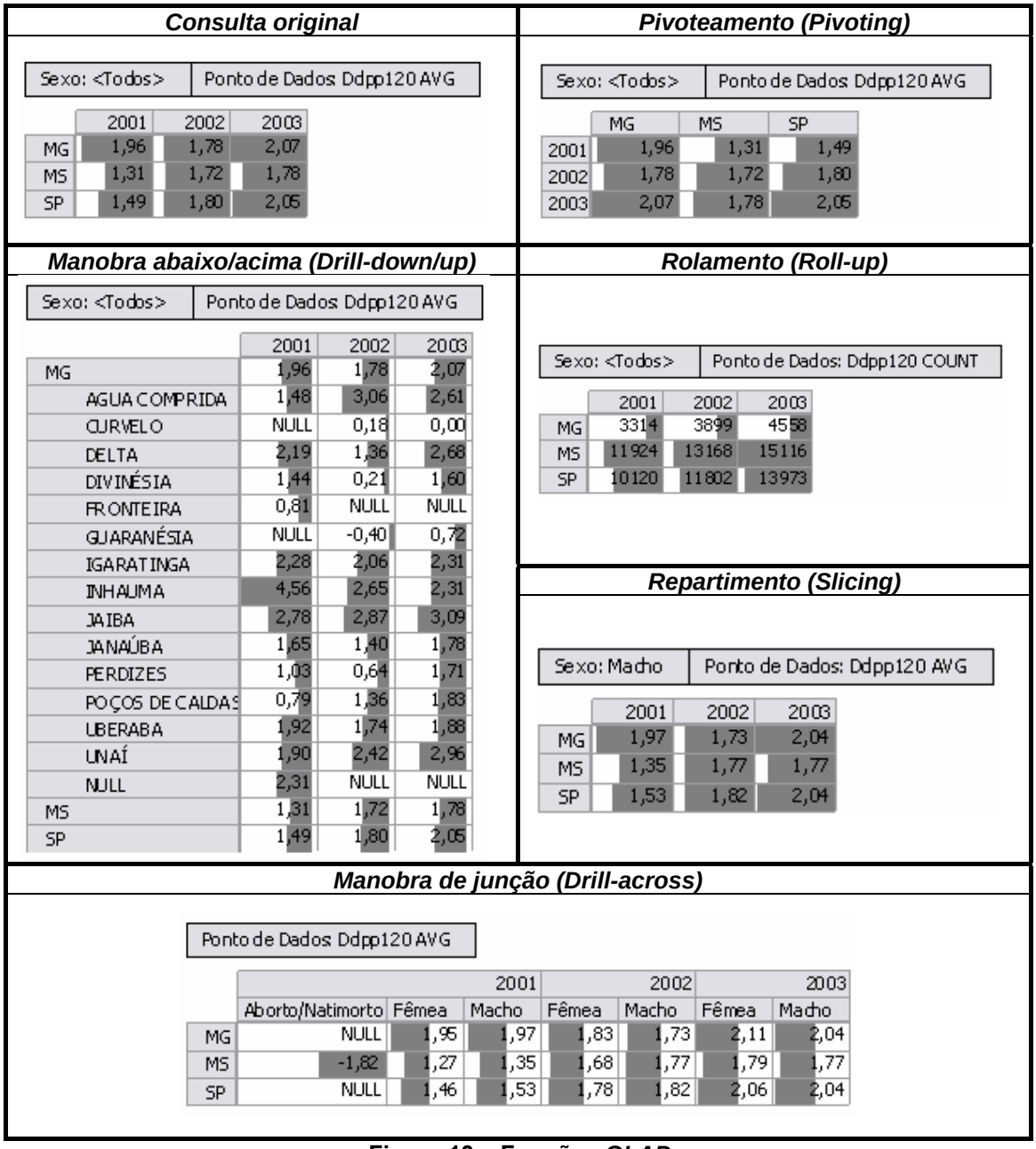


Figura 13 – Funções *OLAP*.

3.4. Mineração de dados

Esta Seção introduz conceitos básicos sobre a mineração de dados (*Data Mining*) e demonstra o porquê se optou pela técnica de visualização de dados.

Descoberta de conhecimento em base de dados (*Knowledge Discovery in Databases – KDD*) foi formalmente definido por Fayyad, Piatetsky-Shapiro e Smith (1996) como “um processo, de várias etapas, não trivial, interativo e iterativo, para a

identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, embutidos nos dados de uma grande base".

Parece não haver consenso entre diversos autores, sobre os termos *KDD* e mineração de dados, alguns consideram sinônimos, outros consideram mineração de dados como a etapa principal (extração de padrões) do processo de *KDD*. Também podem ser encontrados os termos arqueologia de dados, extração de conhecimento, descoberta de informação, coleta de informação e padronização de dados. Enquanto que o termo mineração de dados é comumente utilizado por estatísticos e analistas de dados, os pesquisadores de inteligência artificial preferem utilizar *KDD* (AMARAL, 2001). Neste trabalho foi padronizado chamar de mineração de dados.

Deixando de lado, as divergências conceituais, seguem os benefícios da mineração de dados:

- Encontrar padrões de comportamento, ocorrências incomuns e relacionamentos entre atividades e dados que prometem melhorar a posição dos negócios (INMON; TERDEMAN; IMHOFF, 2001). Ex.: Determinar o perfil genético das progênies oriundas do uso incorreto de biotecnologias reprodutivas;
- Capacidade de transformar impressões em fatos (COREY et al., 2001). Ex.: Demonstrar como o rebanho de animais *comerciais* (dedicados à produção de carne) incorporam o material genético dos rebanhos de *seleção* (dedicados à seleção da raça);
- Permite que processos complexos sejam entendidos e possivelmente alterados, aumentando, assim, a capacidade de predição dos resultados e seus impactos (AMARAL, 2001). Ex.: Entender como e onde erros de seleção e acasalamento ocorrem, para que este quadro seja revertido e o rebanho maximize seu ganho genético;
- Encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório (REZENDE, et al. 2003). Ex.: Conhecer quais rebanhos produzem animais de elevado potencial genético, para utilizar estes, como reprodutores;
- Tornar as pessoas mais racionais no processo decisório (GOLDSHMIDT; PASSOS, 2005). Ex.: Identificar touros subutilizados de elevado potencial genético para serem utilizados no processo reprodutivo.

O processo de mineração de dados possui uma filosofia totalmente diferente das consultas *OLAP*. A consulta *OLAP* automatiza a estatística clássica, ao extrair informações que são apresentadas ao usuário de forma multidimensional, ela é utilizada para estatísticas descritivas e testes de hipótese. No processo de mineração de dados, não há hipóteses formuladas, o usuário extrai padrões ocultos nos dados, mostrando relacionamentos nos mesmos. A Figura 14 mostra estas diferenças e suas relações com a base de dados e *Data Warehouse*, além da natureza do conhecimento humano.

A abordagem mais interessante da mineração de dados em relação à estatística tradicional é que, em mineração de dados não necessita de uma hipótese formulada *a priori*. Isto é muito importante, dado que, o conjunto de dados trabalhados podem ser muito mais volumosos e heterogêneos que os utilizados na estatística tradicional (SHIMABUKURO, 2004).

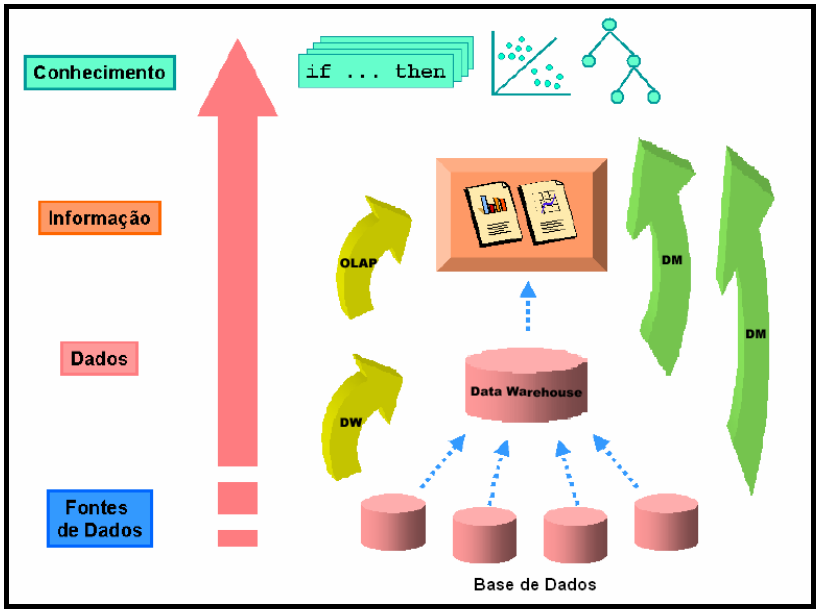


Figura 14 – Relações entre base de dados e *Data Warehouse*; diferenças entre os processos de *Data Warehousing* (*DW*), consulta *OLAP* e mineração de dados (*DM*); natureza do conhecimento humano.
Fonte: Rezende et al. (2003, p. 325).

Como pode ser observado na Figura 14, aplicação de técnicas de mineração de dados pode ser realizada diretamente sobre bases de dados operacionais. Porém, questões como dados espalhados por diversos arquivos ou sistemas, apresentarem diferentes organizações hierárquicas, estarem com valores inesperados ou por ausência de informações, podem inviabilizar a aplicação de mineração de dados, além

de motivar a aplicação de técnicas de integração para um processo eficiente e eficaz (CHEN; HAN; YU, 1996). Quando uma ferramenta de mineração de dados acessa um *Data Warehouse*, os dados já estão limpos e integrados, reduzindo drasticamente o trabalho e tempo da etapa operacional de pré-processamento (GOLDSCHMIDT; PASSOS, 2005; REZENDE et al., 2003).

O *Nelore Business Intelligence* possui dupla aptidão, pois é um repositório de dados que além de proporcionar consultas *OLAP*, também permite a aplicação de técnicas de mineração de dados, pois mantém a granularidade no nível atômico[9]. Deste modo temos uma fonte de dados limpa e integrada que favorece a aplicação de técnicas de mineração de dados.

O processo de mineração de dados depende de três classes de usuários, é composto de cinco etapas, sedo que nas três intermediárias é inevitável o uso de recursos de informática, é interativo e iterativo (Figura 15):

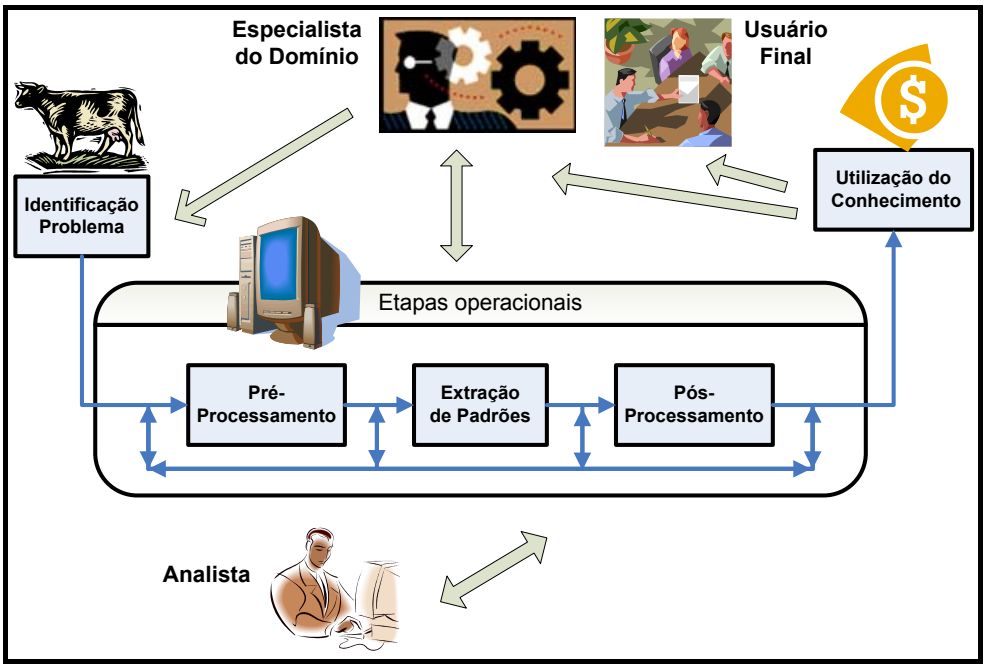


Figura 15 – Etapas e usuários do processo de mineração de dados.

As classes de usuários de um processo de mineração de dados e seus respectivos representantes no PMGRN – Nelore Brasil estão listados à seguir

[9] **Nível atômico:** Menor nível de granularidade de um dado. Ex.: A avaliação genética atribui DEPs aos animais, portanto este é o nível atômico. Já a média da DEP por fazenda é um dado agregado.

(GOLDSCHMIDT; PASSOS, 2005; BATISTA, 2003; REZENDE et al., 2003; AMARAL, 2001). Embora, em muitas situações uma pessoa pode assumir mais de um papel e um papel pode ter vários representantes.

- **Especialista do domínio:** Usuário que possui amplo conhecimento do domínio da aplicação. É o responsável pela identificação do problema e pela tomada de decisão em utilizar os conhecimentos adquiridos. Neste caso, domínio é a avaliação genética e a coordenação do programa de melhoramento genético, especialistas são os pesquisadores do PMGRN – Nelore Brasil;
- **Analista:** Usuário responsável pelas etapas operacionais, possui amplos conhecimentos de informática, dos aplicativos e técnicas de mineração de dados, bem como das etapas que compõem o processo. Também são aqui representados pelos pesquisadores do PMGRN – Nelore Brasil;
- **Usuário Final:** Usuário que se beneficia do conhecimento extraído, embora não precise ter conhecimento profundo do domínio da aplicação e do processo mineração de dados. Pode ser uma pessoa jurídica como uma empresa ou instituição, representados pelos criadores associados e consultores certificados do PMGRN – Nelore Brasil, além da equipe da ANCP.

As etapas de um processo de mineração de dados, exemplificados com dados deste trabalho, estão listados abaixo:

- **Identificação do problema:** Etapa em que o especialista do domínio identifica os problemas dos usuários finais e os objetivos a serem atingidos (REZENDE et al., 2003; AMARAL, 2001). Uma análise cuidadosa deve ser realizada para que se alcance uma melhor compreensão do domínio (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996). O especialista então descreve informalmente o problema e entrega ao analista (BATISTA, 2003). Ex.: Nos últimos anos parece ter ocorrido o uso indiscriminado de biotecnologias reprodutivas como transferência de embriões (TE) e fertilização *in vitro* (FIV), será que os reprodutores foram selecionados corretamente?;
- **Pré-processamento:** Inicia com o analista mapeando os dados relativos ao problema entregue pelo especialista do domínio, havendo grande interação entre representantes destes dois papéis (BATISTA, 2003). É a etapa mais demorada do processo de mineração de dados, chegando a ocupar 80% de todo o tempo (MANNILA, 1997). Compreendem funções relacionadas à capitação, organização, tratamento e preparação dos dados necessários para

a etapa de extração do conhecimento, dividida em oito sub-etapas (GOLDSCHMIDT; PASSOS, 2005):

- **Seleção de dados:** Consiste na identificação das variáveis do banco de dados que devem, efetivamente, participar do processo de mineração de dados e da integração destas num arquivo. Não é necessária a integração, se todas variáveis são coletadas de um *Data Warehouse*. Obs.: Dados extraídos do *Nelore Business Intelligence* via *OLE DB*[10], dispensando integração;
- **Limpeza:** Eliminação ou substituição de valores ausentes ou ruidosos[11]. Esta sub-etapa é dispensável caso a coleta provenha de um *Data Warehouse*. Obs.: Sub-etapa desnecessária, pois os dados foram extraídos do *Nelore Business Intelligence*;
- **Codificação:** Atividade que busca uma melhor representação dos dados, transformando variáveis numéricas em categóricas ou vice versa. Ex.: Foi realizado a codificação numérico-categórica das DEPs analisadas, transformando-as em grupos (TOP 25%, TOP 50% e BOTTON 50%) segundo o percentil;
- **Enriquecimento:** Agregação de mais informações aos registros já coletados. Obs.: Não foi utilizada;
- **Normalização de dados:** Consiste em ajustar a escala de valores de cada atributo em intervalos regulares, como *[-1;+1]* ou *[0;100%]*, evitando que atributos com escalas de valores irregulares influenciem de forma tendenciosa os resultados da mineração de dados. Obs.: Embora cada DEP possua uma escala de valor diferente, esta sub-etapa foi dispensada, dado que, a aplicação de filtros e seleção do conjunto de dados foi realizada nas variáveis codificadas pelos percentis;
- **Construção de atributos:** Geração de novos atributos a partir dos existentes, como criar o número de filhos produzidos por ano, dividindo o número de filhos pela idade do animal. Ex.: Foi construído o atributo

[10] ***OLE DB*****:** *Object Linking and Embedding for Databases* – meio utilizado pela Microsoft para acessar dados armazenados de diferentes formas (Wikipedia, 2006).

[11] **Dados ruidosos:** Dados errados ou que contenham valores considerados divergentes (*outliers*) do padrão esperado.

NF120/Idade, para ter uma idéia de quantas progênies avaliadas um reprodutor deixa por ano e se uma matriz deixa mais de uma, ela com certeza, foi submetida à FIV ou TE;

- **Correção de prevalência:** Consiste em um eventual desequilíbrio na distribuição de registros com determinadas características. Muito utilizado em atividade preditiva de classificação. Suponha que numa base de dados, apenas 0,5% das matrizes fossem submetidas à FIV. Se quiséssemos descobrir um modelo de conhecimento voltado à classificação de novos animais aptos a esta biotecnologia, este poderia ser tendencioso. Obs.: Como trabalhamos com mineração visual de dados, uma atividade descritiva, esta sub-etapa foi dispensada;
- **Partição do conjunto de dados:** Divisão do arquivo em dois conjuntos de dados, treinamento e teste, para processos automáticos de mineração de dados. Obs.: Como trabalhamos com mineração visual de dados, não necessitamos de conjunto de treinamento, portanto esta sub-etapa foi dispensada;

- **Extração de padrões:** Principal etapa do processo de mineração de dados, consiste em aplicar um algoritmo de aprendizado de máquina, uma técnica de visualização ou um teste estatístico na busca efetiva por conhecimentos novos e úteis a partir dos dados (GOLDSCHMIDT; PASSOS, 2005; RAZENTE, 2004; AMARAL, 2001; MONARD et al., 1997; FAYYAD; PIATETSKY-SHAPIRO; SMITH et al., 1996). Como não existe um único processo ou algoritmo ideal para todas as aplicações, pode ser realizado a escolha de vários deles (REZENDE, et al. 2003; BREIMAN, 1996; WOLPERT, 1992). Ex.: Foram utilizadas técnicas de visualização de Mapa de Perfil (*Profile Chart*) e Gráfico de Dispersão (*Scatter Plot*);
- **Pós-processamento:** Etapa que envolve a visualização, análise e interpretação do modelo de conhecimento gerado pela etapa da extração de padrões, para que cheguem ao usuário final apenas padrões relevantes (GOLDSCHMIDT; PASSOS, 2005; REZENDE et al., 2003). Os conhecimentos extraídos devem apresentar compreensibilidade[12] e interassibilidade[13]. As

[12] **Compreensibilidade:** Estimação da simplicidade do modelo relacionado com sua facilidade de interpretação por um ser humano (CRAVEN; SHAVLIK, 1995);

[13] **Interassibilidade:** Valor de um padrão combinando validade, novidade, utilidade e simplicidade (SILBERSCHATZ; TUZHILIN, 1995).

técnicas de visualização permitem reduzir a quantidade de padrões desinteressantes e a complexidade do processo (GANESH et al., 1996). Ex.: O aplicativo de visualização utilizado nesse trabalho possui uma série de recursos para simplificar e transformar os padrões visuais adquiridos, além de organizar e apresentar resultados de diferentes formas (recursos descritos na Seção 3.5).

O processo de mineração de dados é dito iterativo, pois há possibilidade de repetições parciais ou integrais, de qualquer uma das etapas (Figura 15) na busca de resultados satisfatórios por meio de refinamentos sucessivos. Ele é dito interativo, pois há necessidade de atuação do homem como responsável pelo gerenciamento do processo e recursos computacionais, além da interpretação dos resultados. Técnicas de visualização por serem iterativas e interativas contribuem na exploração dos dados, extração e avaliação do conhecimento descoberto (KEIN, 2001; ROHRER; SIBERT; EBERT, 1999; WONG, 1999).

Sistemas operacionais, *Data Warehouse*, técnicas estatísticas, aprendizado de máquina e técnicas de visualização são elementos de apoio ao processo de mineração de dados. (REZENDE et al., 2003; PIATETSKY-SHAPIRO, 1991). Segundo Mendonça Neto e Sunderhaft (1999) e Mendonça Neto et al. (2000), o processo de mineração de dados apresenta três níveis, quanto à interação do especialista do domínio com a técnica utilizada (Tabela 6). O nível abordado neste trabalho foi mineração visual de dados.

Tabela 6 – Classificação dos níveis de mineração de dados quanto à interação com o especialista do domínio.

Nível	*Definição*	*Utilidade*
Mineração visual de dados	Técnicas que combinam visualização e exploração, possibilitando a interação do especialista do domínio com um repositório de dados. Normalmente permitem a exploração inteligente destes dados por meio de controles dos gráficos e seleção interativa da informação a ser analisada	Ganhar uma visão de alto nível dos dados que estão sendo explorados. Caracterizar as entidades que estão sendo quantificadas por estes dados. Entender como estas entidades se comportam no domínio. Identificar classes, tipos, ou grupos homogêneos de entidades. Ganhar conhecimento básico de como estas entidades relacionam entre si.
Extração automática de padrões	Técnicas que analisam dados para extrair padrões automaticamente. Esses padrões precisam ser analisados pelo especialista do domínio, que irá transformá-los em conhecimento. Englobam técnicas de extração de aglomerações (*clusters*), produção de padrões temporais e descoberta de associações entre variáveis.	Identificar classes, tipos, ou grupos complexos de entidades. Caracterizar e entender relações complexas entre estas entidades.
Construção de modelos	Técnicas que constroem modelos automaticamente a partir de um repositório de dados. O modelo deixa explícito o conhecimento, sem necessitar da análise do especialista do domínio (este apenas analisa sua validade). Englobam técnicas que produzem árvores de classificação, redes neurais e regressão múltipla.	Codificar relações complexas em modelos explícitos, prontos para serem usados para classificação, estimação, e predição dentro do domínio.

Fonte: Extraído de Mendonça Neto e Sunderhaft (1999) e Mendonça Neto et al. (2000).

3.5. Visualização de dados

Esta Seção demonstra fundamentos e importâncias das técnicas de visualização, introduz conceitos da mineração visual de dados e apresenta recursos do aplicativo *Spotfire* (Spotfire, 2000).

A mineração visual de dados integra técnicas de visualização à mineração de dados, em que objetos de uma base de dados podem ser vistos em níveis diferentes de granularidade e abstração, ou em diferentes combinações de dimensões. A visualização pode ser ainda integrada às diversas etapas da extração automática de padrões para facilitar o entendimento do processo e resultados (RAZENTE, 2004). Neste trabalho, foi utilizado o primeiro caso citado, em que o aplicativo de visualização, *Spotfire*, foi utilizado de forma interativa e iterativa no processo de mineração de dados.

Trabalhos da área da psicologia vêm demonstrando que os seres humanos são excelentes em processar cenas visuais e muito ruins em processar dados tabulares, pois nossa memória de curto prazo é pequena, para ser exato, esta memória é de apenas 7±2 itens (MILLER, 1994). Uma pessoa ao tentar encontrar o melhor touro, segundo seus objetivos e critérios de seleção, numa tabela de um Sumário contendo 50 touros, ficaria perdido, pois ao chegar à última linha, teria esquecido, no mínimo, 40 destes animais.

Com as técnicas visualização de dados, o sentido da visão passa a ser explorado, aproveitando as capacidades de percepção e cognição do ser humano (ROHRER; SIBERT; EBERT, 1999; WONG, 1999). Continuando com a ilustração do parágrafo anterior, a mesma pessoa poderia plotar os 50 touros num gráfico, atribuindo diferentes cores e formas às DEPs, ela encontraria visualmente, o melhor touro segundo seus objetivos e critérios de seleção.

Trabalhando com ferramentas de visualizações adequadas, dentre elas, o *Spotfire*, Almeida e Mendonça Neto (2001) e Mendonça Neto et al. (2000) afirmam que essas ferramentas de visualização combinam poderosos recursos visuais com controles extremamente fáceis de operar, permitindo que especialistas do domínio possam explorar dados interativamente, de forma tão eficiente, que eles próprios podem encontrar padrões interessantes de informação útil, sem precisar usar algoritmos automatizados neste processo. Com essas características, pesquisadores da área do melhoramento genético podem encontrar padrões que tragam vantagem competitiva ao processo de seleção e acasalamento dos animais, sem a necessidade de uma equipe de recursos humanos capacitada a construir algoritmos complexos.

O *Spotfire* é um aplicativo de visualização interativa e um sistema de consultas dinâmicas. A arquitetura do *Spotfire* (Figura 16) é baseada no conceito de fixar objetos gráficos complexos (quantos o usuário desejar), oriundos dos objetos do repositório de dados, formando as visualizações. Objetos gráficos podem ser pontos coloridos, seleção de hieróglifos ou grupos de polígonos arbitrários, em duas ou três dimensões. Isto permite que usuários do *Spotfire* importem facilmente tabelas bidimensionais e criem visualizações complexas. Essas visualizações podem ser interativamente consultadas, filtradas e o usuário tem a liberdade de aplicar *zoom* e realizar uma observação panorâmica (AHLBERG, 1996).

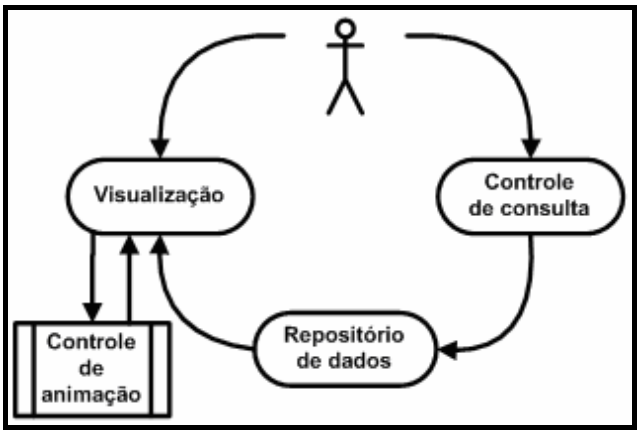


Figura 16 – Arquitetura do *Spotfire.*
Fonte: Ahlberg (1996, p. 25).

Com essa arquitetura (Figura 16), os dados são armazenados num repositório interno do *Spotfire*. Os atributos são armazenados como vetores, com múltiplas indexações para incrementar a performance de várias tarefas de consulta exigidas pelo usuário. A interface com o usuário do *Spotfire* possui dois componentes principais, um controle de consulta com numerosos dispositivos e uma área de visualização, assegurando uma variedade de visualizações pelo controle de animação. Essa arquitetura pode ser observada na tela do aplicativo (Figura 17).

O aplicativo *Spotfire* trabalha com o conceito **Campo Estelar** (***Starfields***) (AHLBERG; SHNEIDERMAN, 1994) que utiliza pontos para representação de registros de dados de várias perspectivas, incluindo ainda, controles de consulta (lado direito da Figura 17) que permitem ao usuário, filtrar o conjunto de dados de interesse por meio de simples seleções com o mouse. As consultas gráficas (lado esquerdo da Figura 17), chamadas de **Consultas Dinâmicas** (***Dynamic Queries***) (SHNEIDERMAN, 1994), criam um nível de interação entre usuário e o computador que é comparável ao de um vídeo game, pois uma seleção no controle de consulta produz sua apresentação, praticamente instantânea, no Monitor do Campo Estelar (*Starfield Display*). O usuário ainda tem à sua disposição, o controle de animação, que proporciona aplicar uma infinidade de *zoom* e observações panorâmicas diretamente no monitor do campo estelar. Segundo Almeida e Mendonça Neto (2001), estes conceitos e propriedades gera um ambiente que permite consultas nebulosas[14] e elimina as inconveniências das consultas que retornem respostas vazias.

[14] **Consultas nebulosas:** Provém de expressões lingüísticas cuja interpretação pode variar de um indivíduo para outro, sendo, portanto, expressões nebulosas (Ex.: “muito”, “médio”, “pouco”) (GOLDSCHMIDT; PASSOS, 2005).

Além do Monitor do Campo Estelar, *o Spotfire* apresenta os componentes:

- **Dispositivo de Consulta (*Query Devices*):** Atributos e seus respectivos valores. O usuário pode selecionar um subconjunto de valores dos atributos para que seja visualizado no Monitor do Campo Estelar;
- **Investigação de Detalhes (*Details on Demand*):** Mostram os valores dos pontos selecionados nas Consultas Dinâmicas;
- **Legenda (*Legend*):** Legenda dos atributos visualizados.

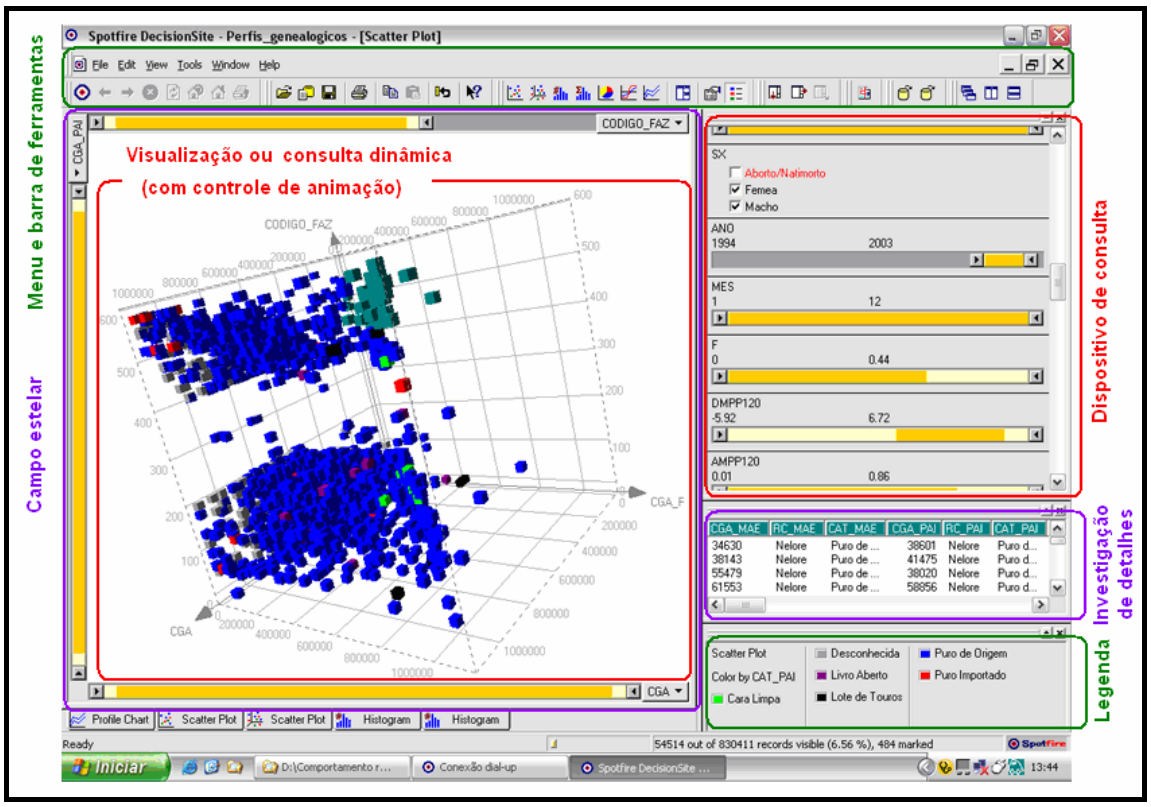


Figura 17 – Tela do *Spotfire* e seus componentes.

O *Spotfire* possui vários recursos para realizar as etapas de pré e pós-processamento, dentre elas citamos algumas utilizadas neste trabalho:

- **Seleção dos dados:** Permite a importação dos dados diretamente de um *SGBD Oracle* via *OLE DB* (Figura 18). Ele apresenta um menu, onde é possível selecionar os atributos, que o especialista do domínio julgar conveniente (Figura 19). Estes recurso foi utilizado para selecionar os atributos do *Nelore Business Intelligence*;

- **Codificação:** Realizada a codificação numérico-categórica para cada uma das DEPs estudadas, dividindo os animais em três grupos, segundo o percentil: TOP 25%, TOP 50% e BOTTON 50% (Figura 20). A codificação no *Spotfire* cria um novo atributo;
- **Construção de atributo:** Permite a criação de um novo atributo, combinando atributos pré-existentes numa equação. Foi criado um novo atributo *NF120/Idade*, indicando o número de filhos deixados no rebanho por ano, por um reprodutor (Figura 21);
- **Pós-processamento:** Gerada uma visualização, o *Spotfire* deixa à disposição do usuário, uma série de ferramentas capazes de trabalhar a imagem no menu Propriedades (*Properties*), como alteração de cores, formas e rotações (Figura 22). Além do menu indicado na Figura 22, é importante lembrar que estamos lidando com uma Consulta Dinâmica, portanto podemos vê-la sobre diferentes perspectivas, apenas com o uso do mouse.

Durante a fase de pós-processamento deste trabalho, o *Spotfire* apresentou as três operações de interação descritas por (SHIMABUKURO, 2004), possibilitando o manuseio das formas de representação:

- **Varredura (*Brushing*):** Permite selecionar e destacar determinados itens de dados, como agrupamentos de interesse, para uma análise mais detalhada;
- **Focalização (*Focusing*):** Permite restringir a análise a determinados intervalos de valores de um atributo;
- **Manuseio do mapa de cores (*Colormap Manipulation*):** Permite configurar esquemas de mapeamento de cores de acordo com as preferências do usuário e objetivos da análise.

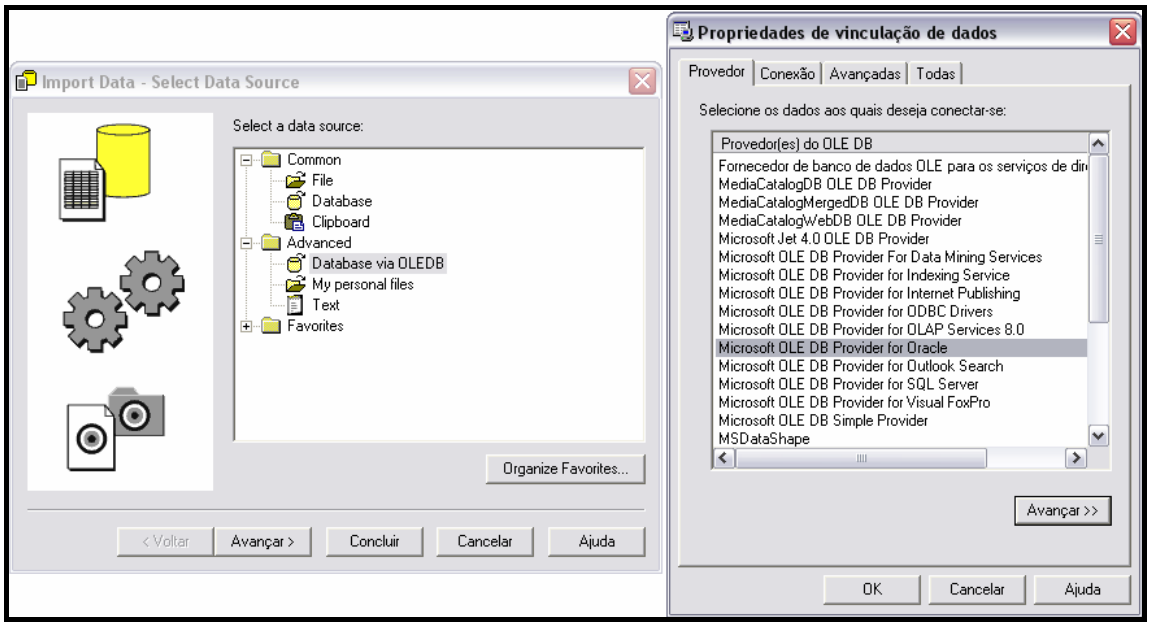


Figura 18 – Importação de dados de um *SGBD Oracle* via *OLE DB*.

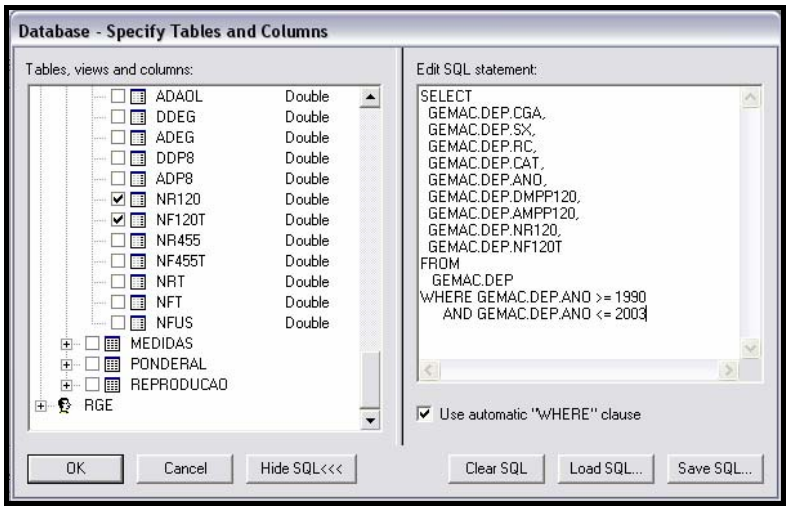


Figura 19 – Seleção de atributos do *Data Warehouse* para visualização no *Spotfire*.

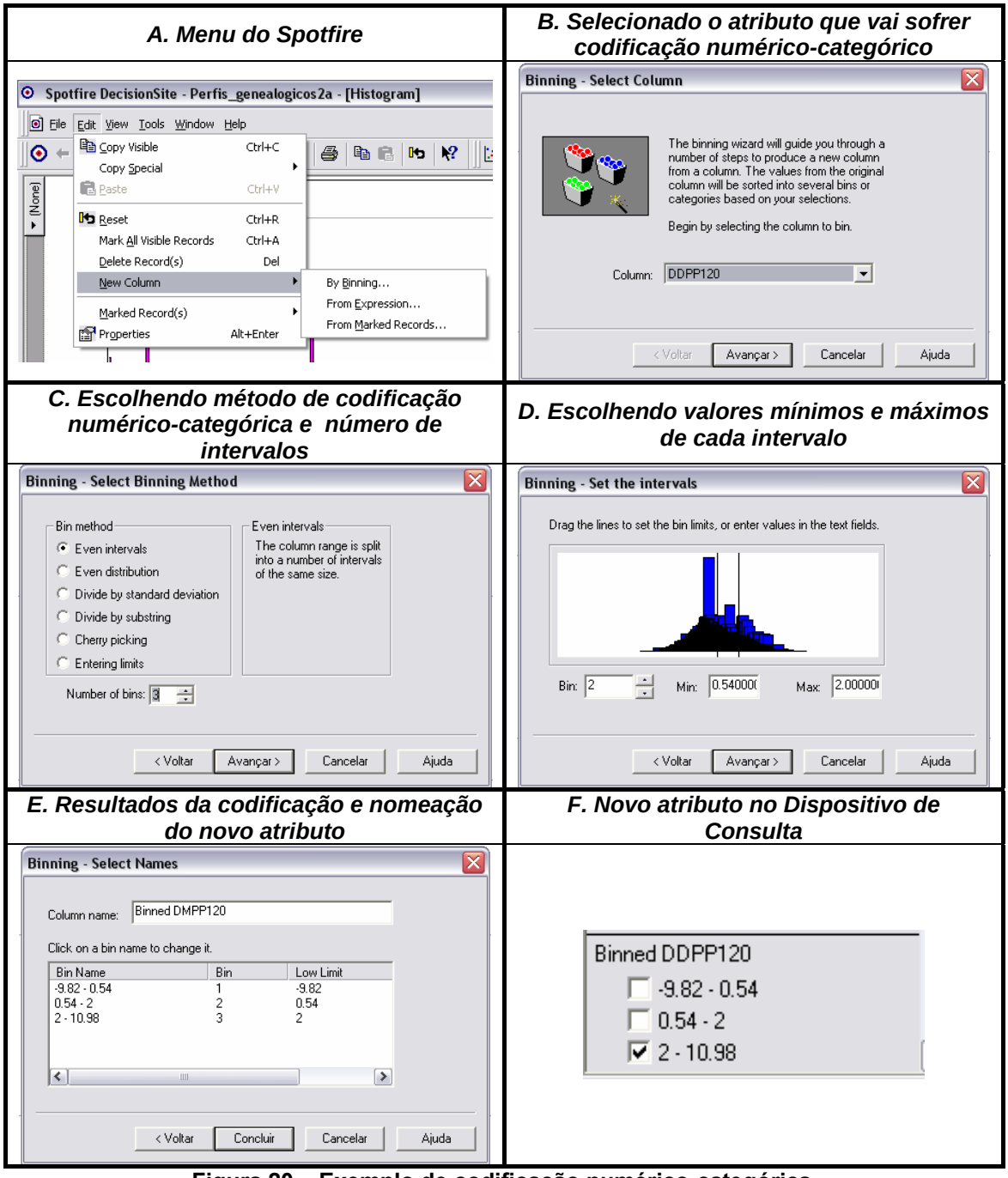


Figura 20 – Exemplo de codificação numérico-categórica.

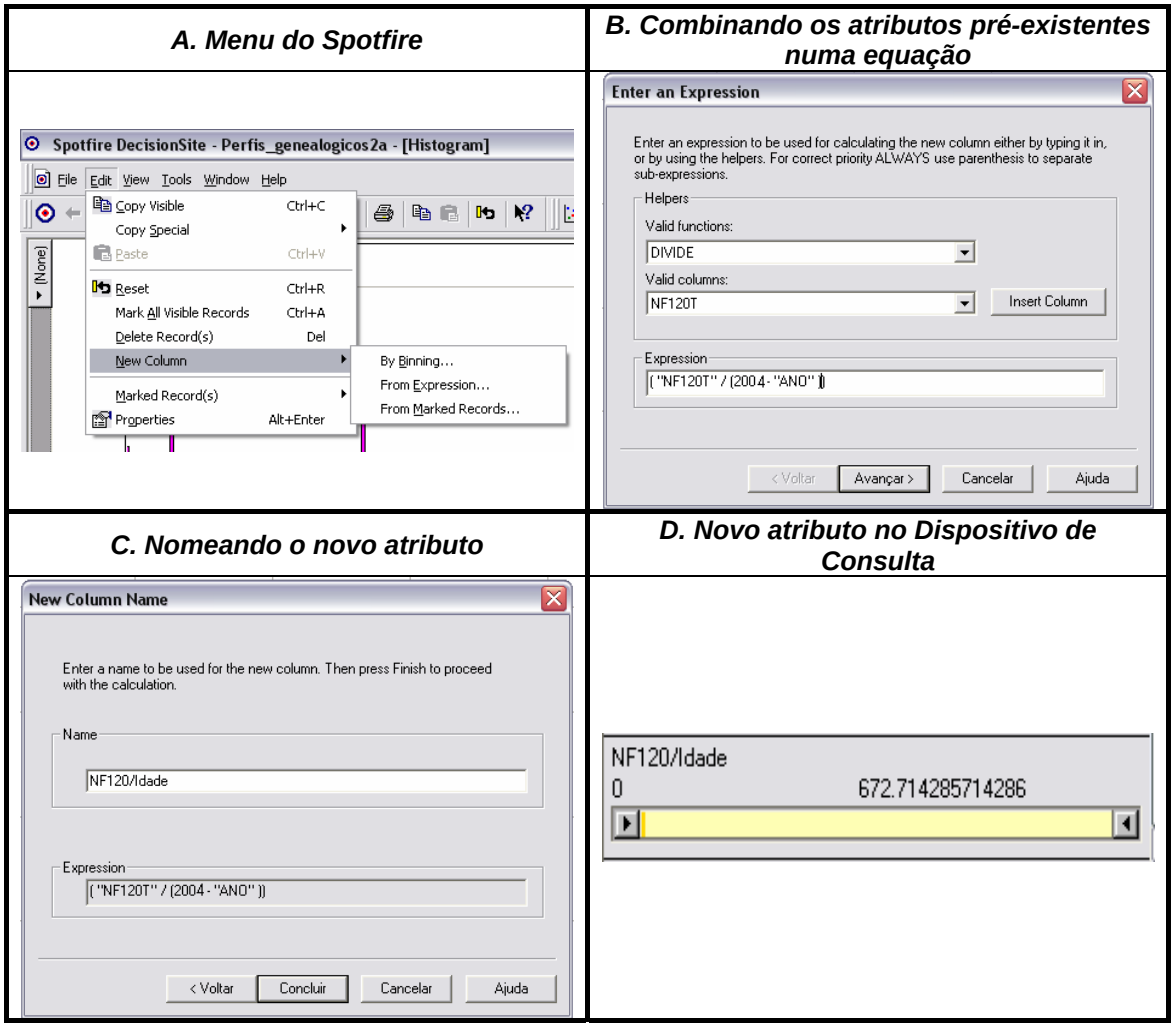


Figura 21 – Exemplo de construção de atributo.

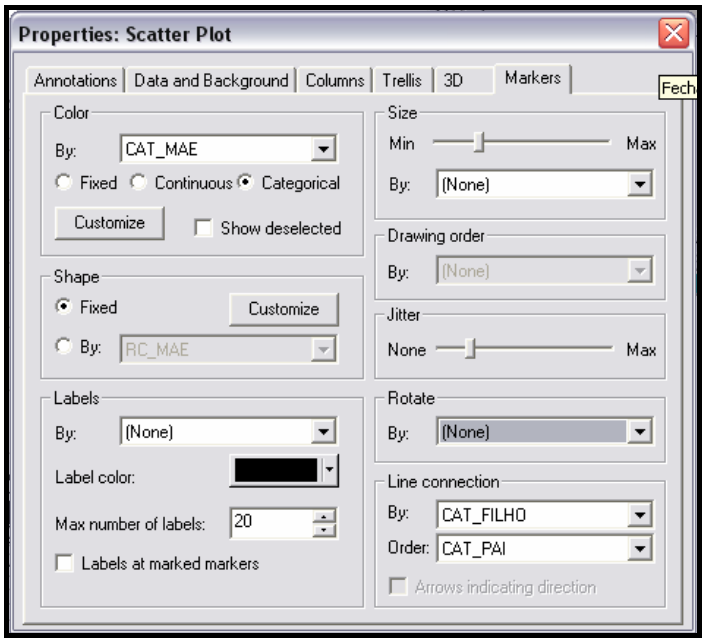


Figura 22 – Exemplo de menu Propriedades.

Segundo Barioni (2002), sempre que é preciso sumarizar grandes quantidades de dados numéricos, utilizam-se gráficos e histogramas, entretanto quando muitas das informações que devem ser apresentadas encontram-se em espaços de altas dimensões, essas técnicas de visualização usuais não são mais indicadas. Keim e Kreigel (1996) descrevem quatro categorias de técnicas de visualização multidimensional: Orientadas a *pixel*, geométricas, iconográficas e hierárquicas (essas duas últimas não foram descritas neste trabalho, porque não são disponibilizadas pelo *Spotfire*). Enquanto que, as técnicas de visualização usuais sumarizam dados, as técnicas de visualização multidimensional conseguem representar graficamente todos os registros de um repositório de dados.

A seguir, são descritas as técnicas de visualização disponíveis no *Spotfire*:

3.5.1. Técnicas de visualização usuais

O *Spotfire* apresenta três formas de representações gráficas que se enquadram às técnicas de visualização usuais: Histograma (*Histogram*), Gráfico de Barras (*Bar Chart*) e Gráfico de Pizza (*Pie Chart*).

Mesmo sendo técnicas de visualização usuais, o *Spotfire*, por ser um aplicativo de mineração visual de dados, permite as três operações de interação, descritas anteriormente. Este diferencial em relação aos aplicativos denominados planilhas eletrônicas, que apresentam gráficos estáticos, conferem aos usuários, a possibilidade de encontrar os padrões implícitos nos dados.

3.5.2. Técnica de visualização multidimensional orientada a *pixel*

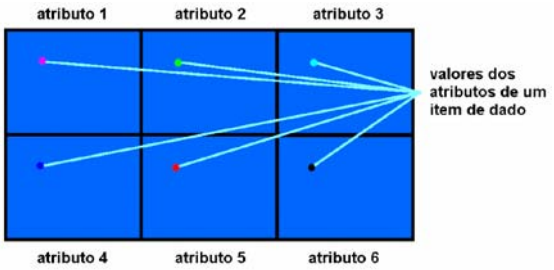
Para cada atributo, os dados são associados a uma posição (janela) na tela e cada valor do atributo é representado por uma cor, deste modo, um único *pixel* é associado para cada dado (Figura 23-A). Com esta técnica, correlação e dependência funcional, entre os atributos, podem ser detectadas pela análise de regiões correspondentes na janela (KEIM, 2000).

Exemplificando, hipoteticamente, como funciona a técnica orientada a *pixel*, a Figura 23-B sugere a correlação entre os atributos *CGA_PAI* e *Percentil* do animal, em

que o touro *CGA_PAI = 200* tende a ter produzir progênies *TOP 25%* (para uma determinada característica hipotética) e o touro *CGA_PAI = 300* tende a produzir progênies *TOP 50%*.

O *Spotfire* possui o Diagrama de Cores (*Heat Map*), que utiliza a técnica orientada a *pixel*. Porém, neste trabalho, não foi utilizada esta técnica, dado que, ao trabalhar com as diferentes amplitudes dos atributos, representados pelas DEPs, e um grande número de dimensões, torna difícil encontrar correlação e dependência entre os atributos em virtude das cores formadas no Diagrama de Cores (Figura 23-C).

A. Idéia básica das técnicas orientadas a pixel. (KEIM, 2000)



B.* Exemplo hipotético de resultados de uma técnica orienta a *pixel.

Atributo 1 CGA	Atributo 2 CGA_PAI	Atributo 3 PERCENTIL
1000	200	Top 25%
1001	300	Top 50%
1002	200	Top 25%
1003	300	Top 50%

C. Diagrama de cores (Heat Map) do Spotfire com dados do PMGRN – Nelore Brasil

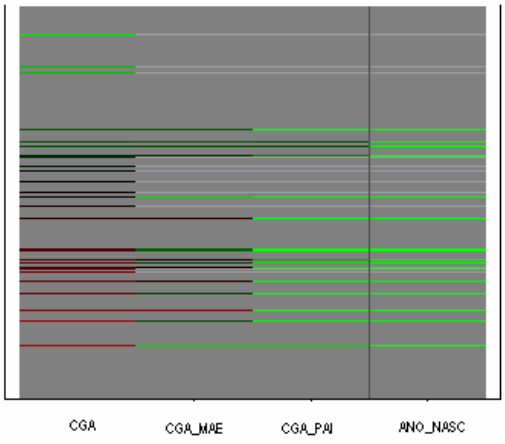


Figura 23 – Técnica de visualização multidimensional orientada a *pixel*.

3.5.3. Técnica de visualização multidimensional de projeção geométrica

Técnicas que projetam em duas ou três dimensões, um conjunto de dados multidimensional, revelando padrões de interesse.

Uma técnica bem usual é denominada técnica de Coordenadas Paralelas (*Parallel Coordinates*), que consiste no mapeamento de uma espaço *K*-dimensional

sobre um gráfico de apresentação bidimensional, usando *k* eixos paralelos e eqüidistantes (INSELBERG; DIMSDALE, 1990). Nesta representação, cada eixo do gráfico representa uma dimensão (atributo) e possui uma escala de valores vertical, compreendendo os valores mínimos e máximos de cada dimensão.

A técnica de coordenadas paralelas transforma relações multivariadas em padrões bidimensionais, permitindo a identificação da distribuição dos dados e correlação dos atributos (INSELBERG, 1997). O *Spotfire* apresenta o Mapa de Perfil (*Profile Chart*) que trabalha com a técnica de coordenadas paralelas (Figura 24).

Como a técnica de coordenadas paralelas e a técnica orientada a *pixel* apresentam funções semelhantes e a primeira demonstrou ser mais eficiente neste estudo, foi eleita para uso.

Outra técnica de visualização multidimensional utilizada foi a de dispersão, denominados no *Spotfire* por Gráfico de Dispersão 2D (duas dimensões – *Scatter Plot 2D*) (Figura 25-A) e Gráfico de Dispersão 3D (três dimensões – *Scatter Plot 3D*) (Figura 25-B). Nessa técnica, cada item de dado é representado por um ponto na área de plotagem do gráfico. O Gráfico de Dispersão 3D permite a rotação do gráfico nas três dimensões, onde o usuário pode explorar todos os seus ângulos.

Segundo (SHIMABUKURO, 2004), uma característica interessante dos Gráfico de Dispersão 3D, é a espacialização de atributos não espaciais, permitindo que uma relação entre itens de dados seja revelada pela proximidade relativa dos mesmos (itens próximos apresentam maior similaridade que itens distantes).

O *Spotfire* permite ainda, a construção de Gráfico de Matrizes (*Trellis Plot*) a partir de qualquer uma das técnicas de visualização multidimensionais disponíveis. O Gráfico de Matrizes (Figura 25-C) é uma matriz de Gráficos de Dispersão (WONG; BERGERON, 1997), onde dados de alta dimensionalidade podem ser representados de forma bidimensional pela projeção dos atributos, organizados aos pares, em forma de uma matriz. Cada célula desta matriz está associada a dois atributos identificados por sua linha e coluna.

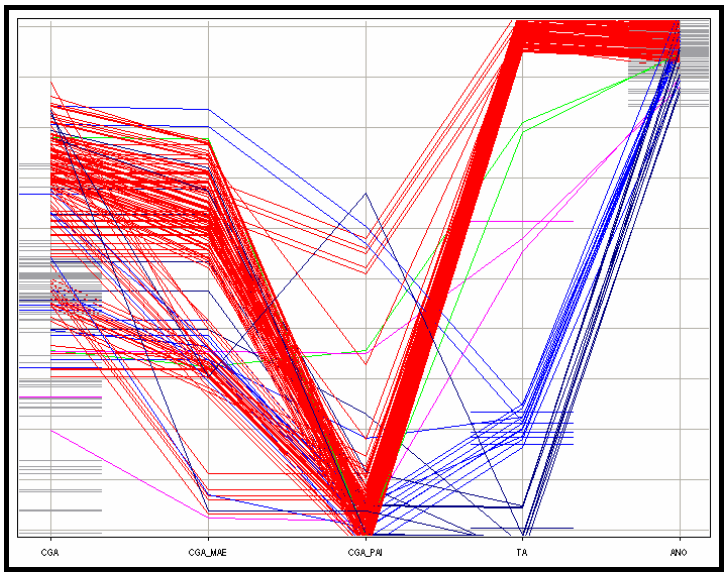


Figura 24 – Técnica de visualização multidimensional de coordenadas paralelas: Mapa de Perfil.

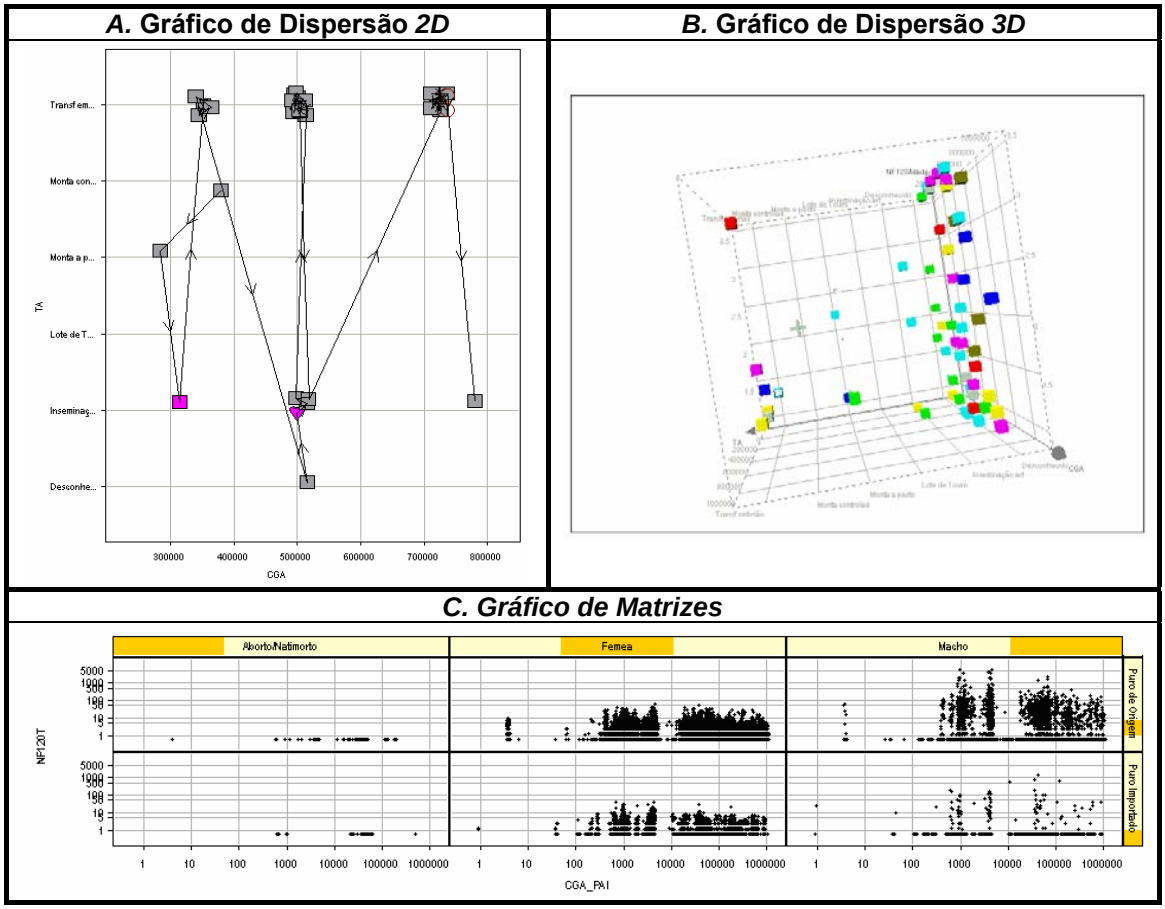
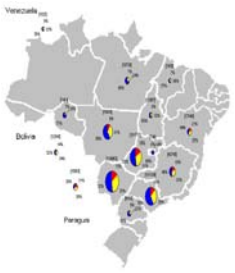
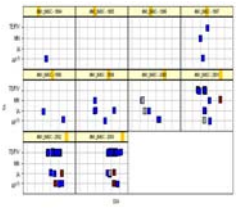
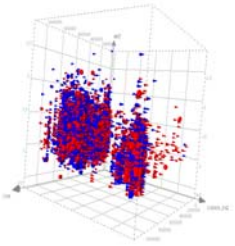
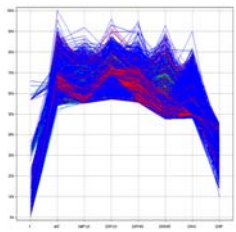


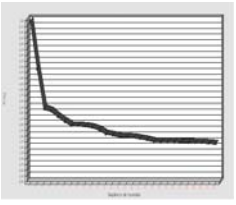
Figura 25 – Técnica de visualização multidimensional de dispersão: A) Gráfico de Dispersão em duas dimensões; B) Gráfico de Dispersão em três dimensões; C) Gráfico de Matrizes.











MATERIAIS E MÉTODOS

4. MATERIAIS E MÉTODOS

Para a presente dissertação foram utilizadas informações pertencentes à avaliação genética de abril de 2005 do PMGRN – Nelore Brasil (LÔBO et al., 2005):

- **Pesagens:** 2.476.111;
- **Aferições de perímetro escrotal:** 399.791;
- **Animais na matriz de parentesco:** 774.870;
- **Animais avaliados:** 278.359;
- **Rebanhos:** 352;
- **Abrangência:** 12 estados brasileiros e mais 3 países (Paraguai, Venezuela, Bolívia);
- **Características avaliadas e disponibilizadas ao criador:** 14 (crescimento: 3; fertilidade: 2; habilidade materna: 1; reprodutivas: 4; probabilidade de permanência no rebanho: 1; quantitativas de carcaça: 3).

As DEPs e acurácias foram estimadas na avaliação genética de 2005, pelos softwares *MTDFREML* (BOLDMAN et al., 1995) e *TKBLUP* (GOLDEN; SNELLING; MALLINCKRODT, 1995).

Nas fazendas participantes do PMGRN – Nelore Brasil são realizadas quatro pesagens anuais dos animais, do nascimento aos 18 meses de idade (janeiro, abril, julho e outubro). As aferições de perímetro escrotal são realizadas duas vezes ao ano, para machos de 9 à 18 meses de idade (janeiro e julho). Algumas fazendas adotam o calendário de pesagens alternativo (fevereiro, março, agosto e novembro). Também são coletadas informações das matrizes e bezerros ao parto e desmama, além da pesagem delas, duas vezes ao ano (abril e outubro).

As informações produtivas, reprodutivas e de genealogia, coletadas nas fazendas são enviadas à ANCP, via correio eletrônico (*e-mail*), passam por uma rigorosa consistência e são incorporadas à base geral de dados (*SisNe*). O *Nelore Business Intelligence* é, então, carregado com os dados oriundos da avaliação genética. O *BIS* serviu como fonte única de dados para as análises aqui utilizadas.

Os softwares utilizados para consultas *OLAP* e mineração visual de dados foram *Oracle Discoverer 4* (Discoverer, 2000b) e *Spotfire* (Spotfire, 2000), respectivamente. O manuseio de arquivos de dados, extraídos do *Nelore Business Intelligence*, e testes estatísticos foram realizados com *SAS* (SAS, 2003). Tabelas e

gráficos clássicos foram construídos com o *Microsoft Excel* (Excel, 2003). As análises foram realizadas no Laboratório de Genética Quantitativa do GEMAC-DG-FMRP-USP, que detém licenças de todos estes softwares.

Foram definidos, arbitrariamente, os termos:

- ***Seleção*:** Animais pertencentes às categorias *Puro de Origem* (*PO*) e *Puro de Origem Importado* (*POI*), responsáveis pela seleção de genes que determinam maior produtividade na pecuária e transmissão destes, aos animais *multiplicadores* e *comerciais*;
- ***Multiplicador*:** Animais pertencentes à categoria *Livro Aberto* (*LA*), responsáveis pela multiplicação dos genes selecionados pelos animais *seleção* e transmissão destes para os animais *comerciais*;
- ***Comercial*:** Animais pertencentes à categoria *Cara Limpa* (*CL*), responsáveis pela produção de carne.

A caracterização da estrutura populacional da Raça Nelore foi realizada com mineração visual de dados. A população analisada compreendeu animais pertencentes às safras de 1990 a 2004, estratificados em três qüinqüênios (*[1990;1994]*, *[1995;1999]* e *[2000;2004]*) e localização geográfica (estados brasileiros e outros países da América Latina). Do grupo de objetos *Reproducao*, foram selecionados os atributos: Identificação do progenitor materno (*CGA_MAE*), variedade (*VARIEDADE_MAE*) e categoria (*CAT_MAE*); identificação do progenitor paterno (*CGA_PAI*), variedade (*VARIEDADE_PAI*) e categoria (*CAT_PAI*); identificação do produto (*CGA_FILHO*), variedade (*VARIEDADE_FILHO*), categoria (*CAT_FILHO*), ano do nascimento (*ANO_NASC*) e sexo (*SX_FILHO*); estado ou país (*UF_FAZ*).

As vias de fluxo gênico para Raça Nelore, assim como a participação das biotecnologias reprodutivas neste processo, foram analisadas com mineração visual de dados. Utilizada mesma população e estratificação da caracterização da estrutura populacional. Do grupo de objetos *Reproducao*, foram selecionados os atributos: Identificação do progenitor materno (*CGA_MAE*) e categoria (*CAT_MAE*); identificação do progenitor paterno (*CGA_PAI*) e categoria (*CAT_PAI*); identificação do animal (*CGA*), categoria (*CAT_FLHO*) e ano de nascimento (*ANO_NASC*); tipo de acasalamento (*TA*).

O progresso genético em função do rebanho (*seleção*, *multiplicador* e *comercial*) foi determinado com consulta *OLAP*, utilizando o grupo de objetos *DEP* (Tabela 7).

Tabela 7 – Atributos e operações utilizadas para análise do progresso genético do MGT.

Tipo do atributo	*Nome do atributo*	*Cálculo*	*Filtros*	*Slicing*
Dimensão	*Ano de nascimento*		= [1990;2004]	
Dimensão	*Categoria*		≠ 'Lote de touros'	
Fato	*MGT*	*AVG*		

O controle da endogamia foi analisado em duas diferentes perspectivas:

- Evolução do coeficiente de endogamia em função da safra e categoria dos animais;
- Identificação do coeficiente de endogamia médio por fazenda.

Foram selecionados animais pertencentes às safras 1994 a 2003, pois assim, têm animais com potencial e idade para serem os reprodutores da atualidade. Foi imposta a restrição de analisar apenas fazendas com no mínimo 100 animais por safra com dados válidos para F. O estudo da endogamia foi realizado com consulta *OLAP*, utilizando o grupo de objetos *DEP* (Tabela 8).

Tabela 8 – Atributos e operações usados para análise do coeficiente de endogamia (F).

Tipo do atributo	*Nome do atributo*	*Cálculo*	*Filtros*	*Slicing*
Dimensão	*Categoria*		= 'Puro de Origem' ou 'Puro Importado'	
Dimensão	*Ano do Nascimento*		= [1994;2003]	2003(1)
Dimensão	*Fazenda*			
Fato	*MGT*	*Count*	≥ 100	
Fato	*F*	*AVG*		

(1) Apenas na análise da endogamia por fazenda. As fazendas foram classificadas em ordem decrescente para F.

Para identificação de padrões do processo de seleção e acasalamento dos animais, foi escolhido, arbitrariamente, o seguinte conjunto de DEPs para o estudo: efeito maternal para peso aos 120 dias (MP120), efeito direto para peso aos 120 dias (DP120) e 450 dias (DP450), perímetro escrotal aos 450 dias (DPE450), idade ao primeiro parto (DIPP) e produtividade acumulada (DPAC). Dentro deste conjunto de DEPs, os animais foram divididos, arbitrariamente, em três classes, pelo percentil (Tabela 9):

- **TOP 25%:** Para as seis características concomitantemente, animais geneticamente superiores;

- **TOP 50%:** Para as seis características concomitantemente, animais geneticamente intermediários;
- **BOTTON 50%:** para as seis características concomitantemente, animais geneticamente inferiores.

Tabela 9 – Codificação numérico-categórica das DEPs.

DEP	*Intervalos de classe*[(1)]		
	TOP 25%	*TOP 50%*	*BOTTON 50%*[(1)]
DP120	[2,00;10,98]	[0,54;2,00[	[-9,82;0,54[
DP450	[6,25;36,06]	[2,35;6,25[	[-30,93;2,35[
DPE450	[0,15;2,64]	[-0,05;0,15[	[-2,09;-0,05[
MP120	[1,20;6,72]	[0,44;1,20[	[-5,92;0,44[
DPAC	[2,38;15,6]	[1,14;2,38[	[-9,74;1,14[
DIPP	[-2,02;-0,38]	]-0,38;-0,17]	]-0,17;1,58]

(1) Subconjunto de valores selecionados com o componente Dispositivo de Consulta, segundo valores descritos por Lôbo et al. (2005).

Como eram necessários conjuntos de atributos provenientes de dois grupos de objetos distintos do *PMGRN-DM*, o *Nelore Business Intelligence* foi acessado pelo aplicativo *Spotfire*, via *OLE DB*, duas vezes, sendo exportadas as tabelas para formato texto (.TXT):

- **Grupo de objetos *DEP*:** Identificação do animal (*CGA*), sexo (*SX*), ano do nascimento (*ANO*) e mês do nascimento (*MES*); coeficiente de endogamia (*F*); DEPs para MP120 (*DMPP120*), DP120 (*DDPP120*), DP450 (*DDPP455*), DPE450 (*DDPE455*), DIPP (*DDIPP*) e DPAC (*DDPAC*); Mérito Genético Total (*MGT*); número de filhos avaliados para DP120 (*NF120T*) e número de rebanhos com filhos avaliados para DP120 (*NR120*); última situação do animal (*USA*). Foram selecionados todos os animais avaliados pelo PMGRN – Nelore Brasil, formando uma matriz de parentesco para os animais analisados;
- **Grupo de objetos *Reproducao*:** Identificação do progenitor materno (*CGA_MAE*) e categoria (*CAT_MAE*); identificação do progenitor paterno (*CGA_PAI*) e categoria (*CAT_PAI*), identificação da progênie (*CGA_FILHO*[15]), raça (*RC_FILHO*), categoria (*CAT_FILHO*), sexo (*SX_FILHO*) e ano de nascimento (*ANO_NASC*); código da fazenda (*CODIGO_FAZ*) e estado (*UF_FAZ*). Foram selecionados animais pertencentes às safras 1994 a 2003,

[15] ***CGA_FILHO*:** Chave substituta para *CGA* na tabela DEPs.

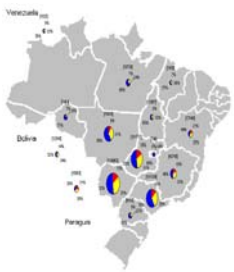
pois assim, têm animais com potencial e idade para serem os reprodutores da atualidade.

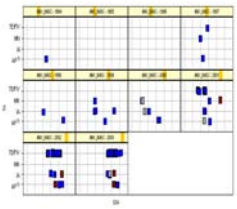
Foram excluídos da análise, animais sem avaliação genética e casos de aborto e natimortos.

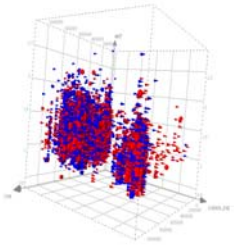
Como uma matriz submetida a TE ou FIV é capaz de produzir mais de um bezerro por ano, foi construído o atributo *NF120/Idade*, utilizando a ferramenta *New Column – From Expression*, conforme a fórmula abaixo:

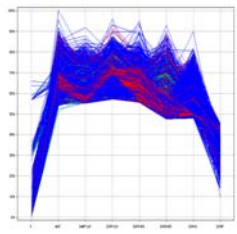
$$NF120/Idade = \frac{NF120}{2004 - ANO\_NASC}$$

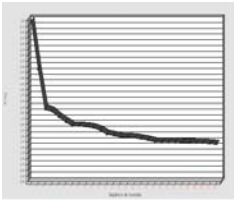
Os coeficientes de correlação de Pearson foram estimados entre todas essas DEPs pelo Procedimento CORR (SAS, 2003), para todo arquivo de dados e para o conjunto de animais TOP 25%. Para extração dos padrões de seleção e acasalamento foi utilizada mineração visual de dados.











RESULTADOS E DISCUSSÃO

5. RESULTADOS E DISCUSSÃO

5.1. Caracterização da estrutura populacional da Raça Nelore

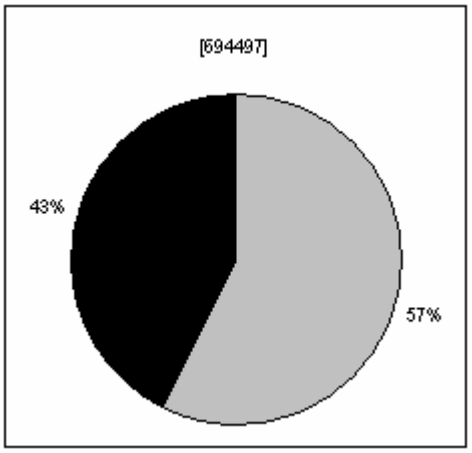
Quanto ao sexo, na população de animais avaliados pelo PMGRN – Nelore Brasil, o número de fêmeas é maior que o de machos em 14 pontos percentuais (Figura 26). Valor esperado, dado que o número de touros necessários num rebanho é menor que o de matrizes, portanto a pressão de seleção e descarte de machos são maiores.

Na análise por categoria (Figura 27), maior discrepância entre sexos encontra-se nos rebanhos *LA* (34 pontos percentuais) e *CL* (22 pontos percentuais), com maior equilíbrio nos rebanhos *PO* (8 pontos percentuais) e *POI* (6 pontos percentuais). Como no rebanho *comercial* (dedicado à produção de carne) e, muitas vezes ocorre no rebanho *multiplicador*, a finalidade do macho é o abate, os criadores tendem a utilizar touros *seleção* ou IA na vacada (visando aumentar a produtividade), portanto o objetivo deles é a seleção de fêmeas, que serão novilhas de reposição do plantel.

A distribuição de sexos por estados e países é bem irregular (Figura 28), o maior desequilíbrio foi constatado em MA (57 pontos percentuais) e o maior equilíbrio, na BA (4,8 pontos percentuais), sendo o único estado onde o número de machos supera ao de fêmeas. Há uma tendência de equilíbrio nos estados tradicionais na exportação de genética, como SP e MG, e desequilíbrio maior nos estados exportadores de carne, como RO e MT. Sugerindo o trânsito de touros jovens criados nos estados dedicados à produção de genética para os produtores de carne, ou seja, touros jovens *seleção* para uso em vacada *multiplicadora* e *comercial*.

Ao longo dos anos, vem aumentando o equilíbrio entre sexos (Figura 29), o primeiro qüinqüênio apresentou a maior discrepância do número de fêmeas em detrimento a machos (34 pontos percentuais) e o terceiro, o maior equilíbrio (2 pontos percentuais). Esses resultados indicam o amadurecimento do PMGRN – Nelore Brasil, dado que, novos rebanhos aderem ao programa de melhoramento genético inscrevendo, principalmente, suas matrizes, levando a grande discrepância entre sexos no primeiro qüinqüênio. Como as progênies de matrizes avaliadas tendem a ter

avaliação genética, devido a matriz de parentesco, o número de machos e fêmeas, no último qüinqüênio, tendeu ao equilíbrio.



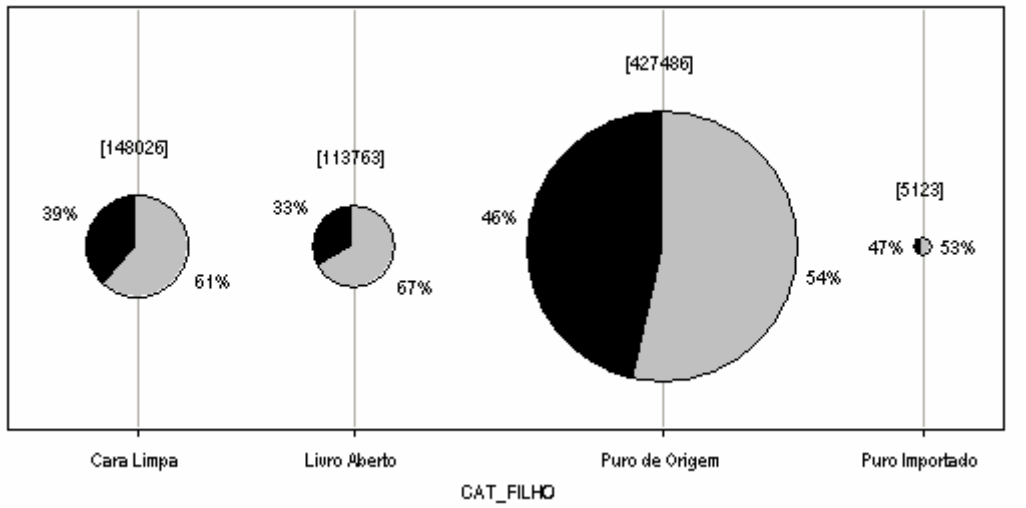
Legenda:

Cor por sexo:
··· Fêmea,
■ Macho;

Tamanho da pizza por número de animais;

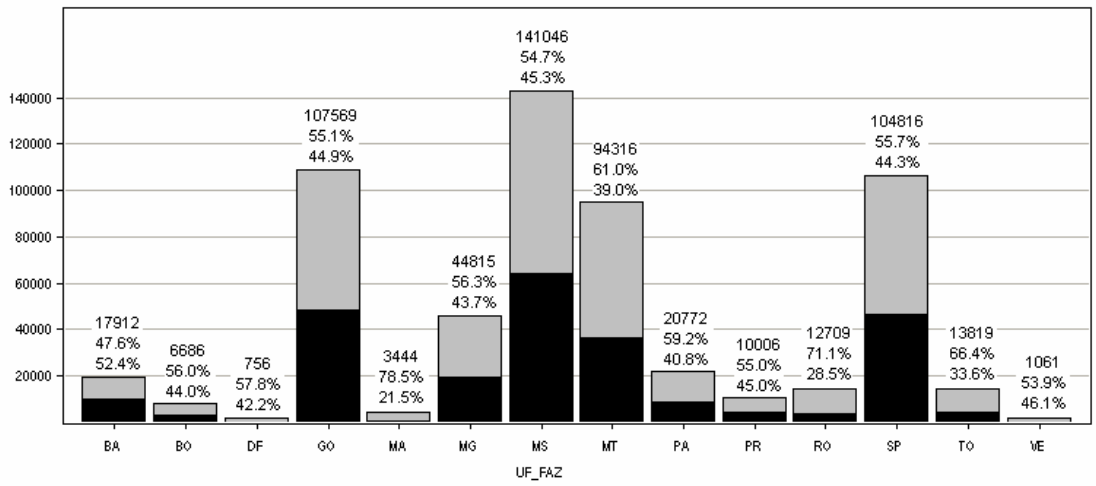
Rótulos: número de animais e porcentagem do setor.

Figura 26 – Distribuição dos animais por sexo.



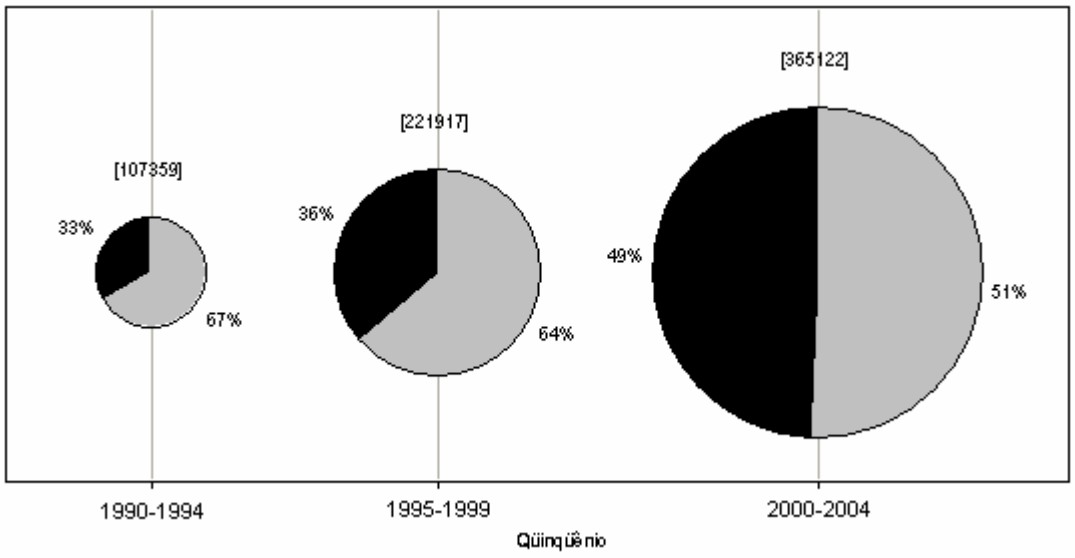
Legenda:
Cor por sexo: ··· Fêmea, ■ Macho;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

Figura 27 – Distribuição dos animais por sexo e categoria.



Legenda:
Cor por sexo: ··· Fêmea, ■ Macho;
Tamanho da barra por número de animais;
Rótulos: número de animais, porcentagem de fêmeas e porcentagem de machos.

Figura 28 – Distribuição dos animais por sexo, estados brasileiros e outros países.



Legenda:
Cor por sexo: ··· Fêmea, ■ Macho;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

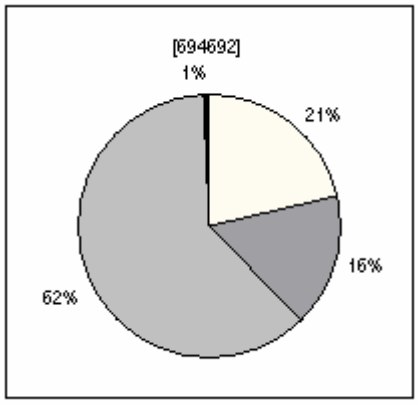
Figura 29 – Distribuição dos animais por sexo e qüinqüênios.

Quanto à categoria dos animais avaliados pelo PMGRN – Nelore Brasil (Figura 30), a maioria são *PO* (62%), seguidos de *CL* (21%), *LA* (16%) e *POI* (1%). Esses números quebram o mito da incidência do programa de melhoramento genético somente nos rebanhos *seleção*, dado que 37% dos animais avaliados pertencem aos rebanhos *multiplicador* e *comercial*.

A distribuição geográfica da categoria dos animais é muito irregular (Figura 31), sendo que há predomínio do rebanho *seleção* em oito regiões (BA, BO, DF, MG, PR, SP, VE e TO), predomínio dos rebanhos *multiplicador* e *comercial* em duas regiões (MA e RO) e equilíbrio em quatro regiões (GO, MS, MT, PA). Contrastando com a distribuição geográfica do sexo (Figura 28), existe uma tendência de haver maior proporção de fêmeas ou equilíbrio em regiões com predomínio de rebanhos *multiplicador* e *comercial*, ou seja, matrizes dedicadas à produção de novilhas de reposição e de carne, respectivamente.

A distribuição temporal da categoria (Figura 32) indica que no primeiro qüinqüênio, o PMGRN – Nelore Brasil atraía poucos rebanhos *multiplicadores* e *comerciais* (29% de participação, somando o número de animais das categorias *CL* e *LA*). Essa proporção aumentou drasticamente no final do século XX, segundo qüinqüênio (categorias *CL* e *LA* passaram a contribuir com 39% dos animais avaliados) estabilizando-se no terceiro qüinqüênio (categorias *CL* e *LA* continuam contribuindo com 39% dos animais avaliados). Este aumento de participação sugere que, este tipo de criador descobriu os benefícios do programa de melhoramento genético para aumento da produtividade de seus rebanhos.

Com a estabilidade da proporção entre rebanhos *multiplicador-comercial* e *seleção* nos dois últimos qüinqüênios (Figura 32), a categoria *LA* aumentou sua participação entre o segundo e o terceiro qüinqüênio (4 pontos percentuais). Esse resultado sugere o uso de touros *PO* em matrizes *CL*, para formação do rebanho *multiplicador*. Quando contrastado à distribuição geográfica da categoria dos animais (Figura 31), temos que, estados com predominância dos rebanhos *multiplicador* e *comercial* ou equilíbrio, são nichos de mercados (MA, RO, MT, PA, GO e MS) para venda de touros jovens *PO*.



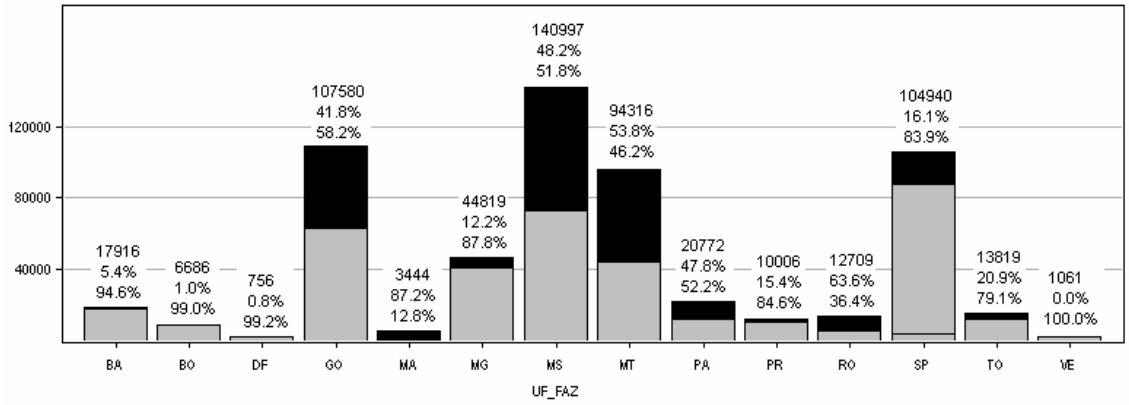
Legenda:

Cor por categoria:
··· *CL*,
··· *LA*,
··· *PO*,
POI;

Tamanho da pizza por número de animais;

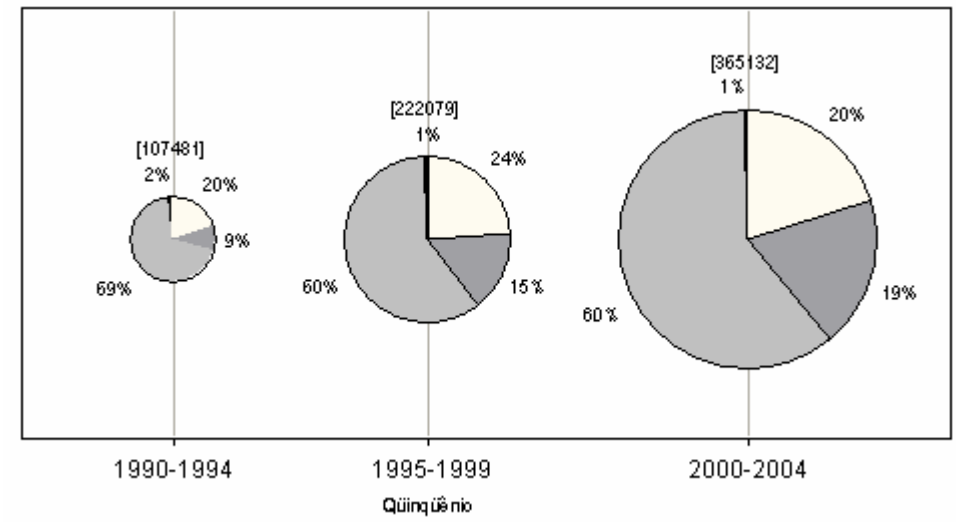
Rótulos: número de animais e porcentagem do setor.

Figura 30 – Distribuição dos animais por categoria.



Legenda:
Cor por categoria: ■ *CL* e *LA* (*comercial* e *multiplicador*), ··· *PO* e *POI* (*seleção*);
Tamanho da barra por número de animais;
Rótulos: número de animais, percentagem das categorias *CL* e *LA*, porcentagem das categorias *PO* e *POI* .

Figura 31 – Distribuição dos animais por classes de categorias, estados e outros países.



Legenda:
Cor por categoria: ··· *CL*, ··· *LA*, ··· *PO*, ■ *POI*;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

Figura 32 – Distribuição dos animais por categoria e qüinqüênios.

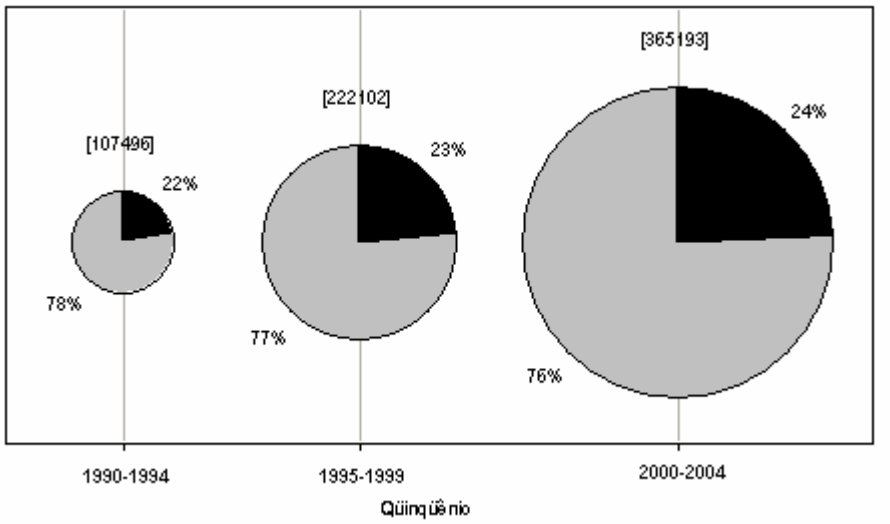
Quanto à variedade (mocho ou padrão) dos animais avaliados pelo PMGRN – Nelore Brasil (Figura 33), a maioria são padrões (76% de média ponderada entre os três qüinqüênios), com ligeira tendência de aumento da participação do mocho (2 pontos percentuais), ao longo dos qüinqüênios.

Embora o mais comum seja o acasalamento entre a mesma variedade, é mais freqüente o uso de touros mochos em matrizes padrões, que o contrário (Figura 34), reportando a história da formação da variedade mocha no Nelore, com uso de touros responsáveis pela difusão dessa variabilidade genética.

A comparação entre variedades e categorias (Figura 35) indica maior participação da variedade mocha nas categorias *LA* (31%) e *PO* (27%) em detrimento a *CL* (9%). A grande participação da variedade mocha na categoria *LA* quando contrastada com acasalamentos entre diferentes variedades (Figura 34), sugere o uso de touros mochos *PO* em matrizes padrões *CL*. Estes resultados são indicativos de nicho de mercado para touros jovens mochos e sêmen.

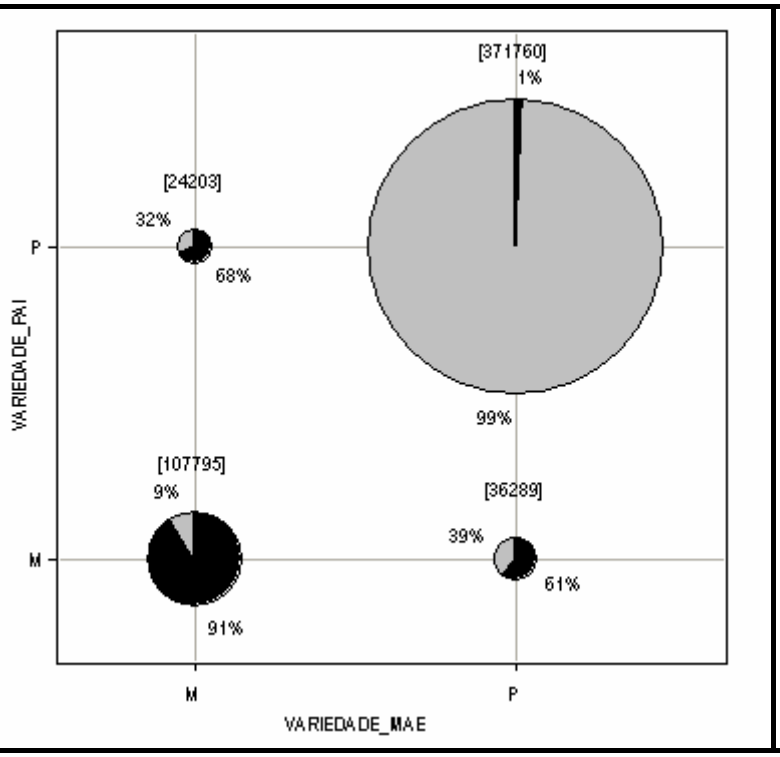
Ao analisar o comportamento temporal para distribuição geográfica da variedade dos animais (Figura 36), detecta-se comportamento heterogêneo. O contraste com as distribuições regionais por categorias (Figura 31) não indica existência de relações concretas sobre a preferência pela variedade, devendo ser em virtude de fatores culturais ou campanhas de marketing das fazendas participantes do PMGRN – Nelore Brasil de cada região.

A análise do crescimento da participação das variedades por qüinqüênios (Figura 37), indica crescimento do mocho (1,8 pontos percentuais) nos animais avaliados pelo PMGRN – Nelore Brasil. Dividindo em regiões, a participação do mocho é crescente em cinco delas (VE, SP, PA, MA e GO), a participação do padrão é crescente noutras cinco (RO, TO, PR, MS e MG), o comportamento é oscilatório em três regiões (MT, BO, BA) e estável numa região (DF). Este comportamento indica nichos interessantes de mercado para vendas de touros jovens, semens, matrizes e embriões para cada uma das variedades.



Legenda:
Cor por variedade:
Mocho,
Padrão;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

Figura 33 – Distribuição dos animais por variedade.



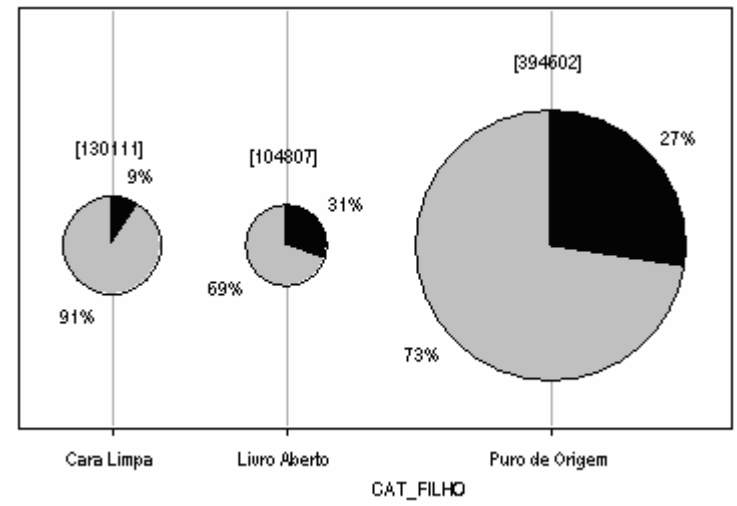
Legenda:

Cor por variedade:
Mocho (M),
Padrão (P);

Tamanho da pizza por número de animais;

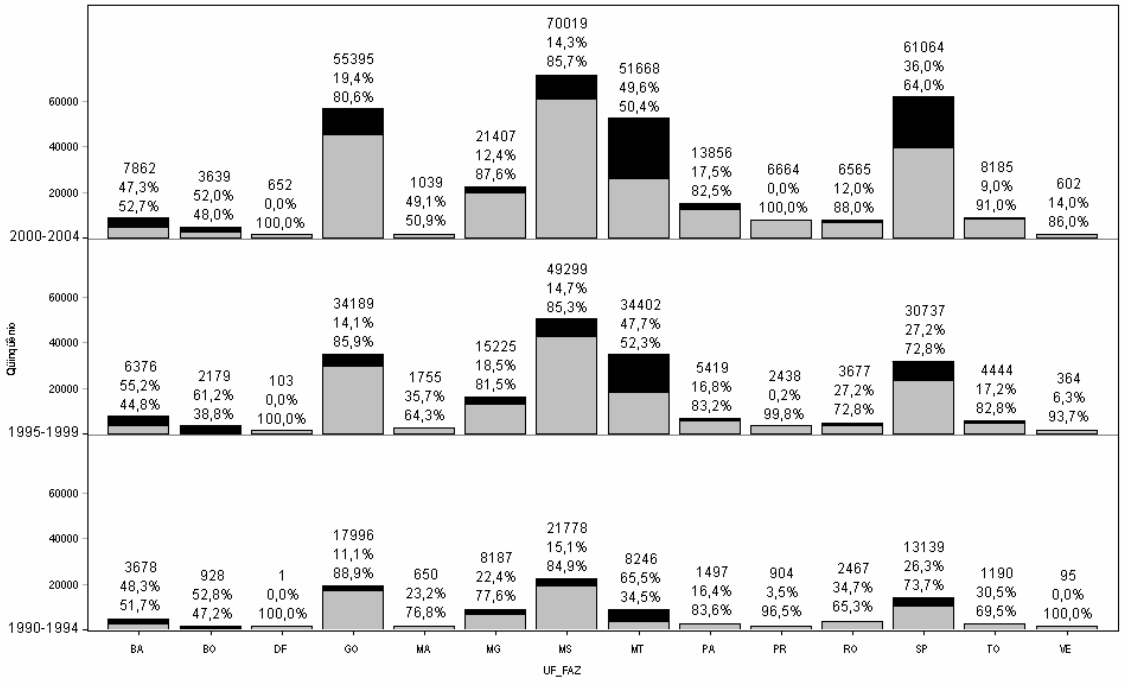
Rótulos: número de animais e porcentagem do setor.

Figura 34 – Distribuição dos animais por variedade da progênie e progenitores.



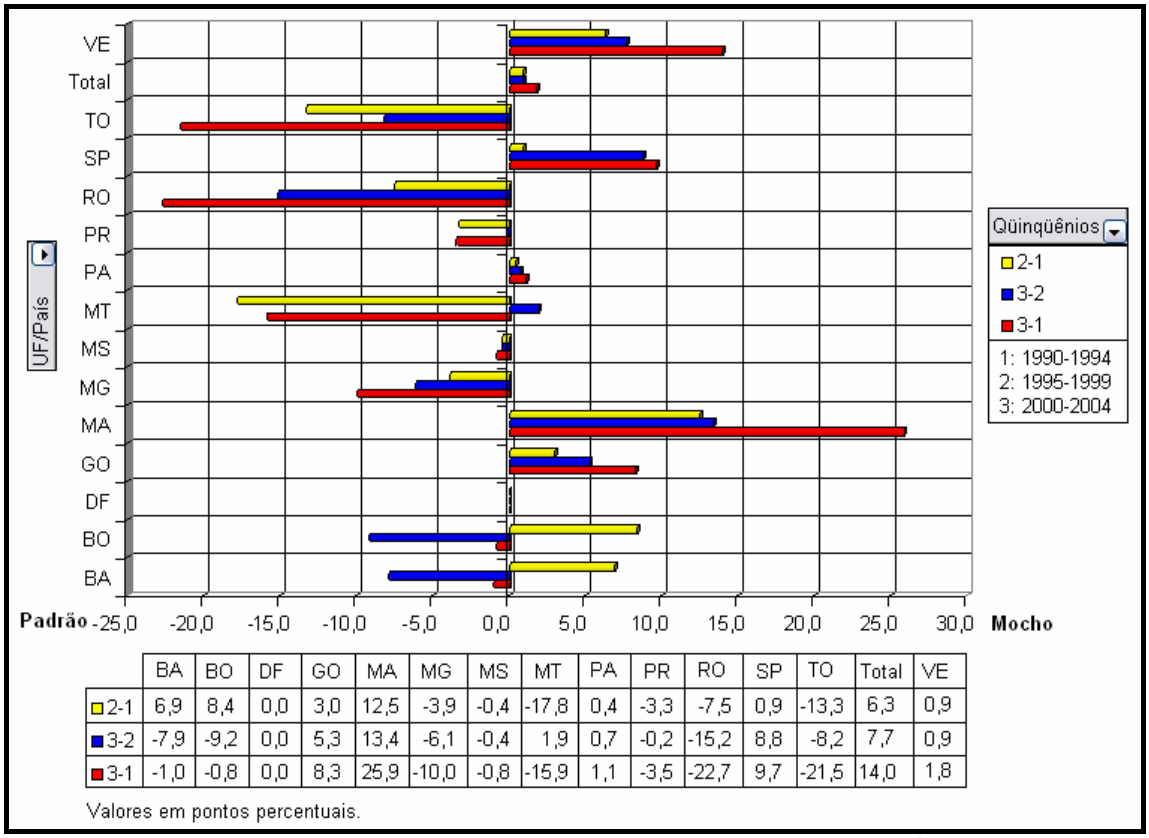
Legenda:
Cor por variedade: ■ Mocho, ··· Padrão;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

Figura 35 – Distribuição dos animais por variedade e categoria.



Legenda:
Cor por variedade: ■ Mocho, ··· Padrão;
Tamanho da barra por número de animais;
Rótulos: número de animais, porcentagem do mocho e porcentagem do padrão.

Figura 36 – Distribuição dos animais por variedade, estados brasileiros e outros países e qüinqüênios.



	BA	BO	DF	GO	MA	MG	MS	MT	PA	PR	RO	SP	TO	Total	VE
2-1	6,9	8,4	0,0	3,0	12,5	-3,9	-0,4	-17,8	0,4	-3,3	-7,5	0,9	-13,3	6,3	0,9
3-2	-7,9	-9,2	0,0	5,3	13,4	-6,1	-0,4	1,9	0,7	-0,2	-15,2	8,8	-8,2	7,7	0,9
3-1	-1,0	-0,8	0,0	8,3	25,9	-10,0	-0,8	-15,9	1,1	-3,5	-22,7	9,7	-21,5	14,0	1,8

Valores em pontos percentuais.

Figura 37 – Crescimento da participação das variedades por qüinqüênios (valores positivos no gráfico indicam crescimento da participação da variedade mocha e valores negativos, da variedade padrão).

5.2. Fluxo gênico na Raça Nelore

Dentro do ciclo produtivo da pecuária de corte, o fluxo gênico deve partir dos restritos rebanhos *seleção* (melhoradores) e atingir os rebanhos *comerciais*, responsáveis pela produção de carne que chega ao varejo (Figura 38). No Brasil, dada às vastas extensões territoriais das fazendas, o rebanho *multiplicador* se confunde tanto ao rebanho de *seleção* (fornecer genética) quanto ao *comercial* (produzir carne). Segundo Lima (2004), será um trabalho patriótico do melhorista quando ele conseguir atingir o efeito do melhoramento animal em todo o segmento da cadeia produtiva, sem exceção, só assim o invernista poderá usar toda a tecnologia existente para melhor suprir o mercado de carne para o consumo interno e externo.

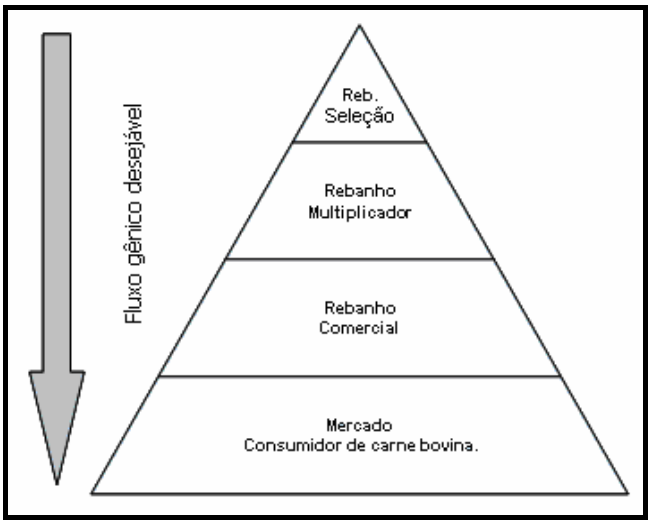


Figura 38 – Fluxo gênico na pecuária de corte.

A princípio, parece que a avaliação genética incide apenas sobre rebanhos *seleção*. Este é um grande equívoco, dado que, um programa de melhoramento genético atinge todo rebanho das fazendas participantes. Portanto, os animais são avaliados conjuntamente, quer seja, *seleção*, *multiplicador* ou *comercial*.

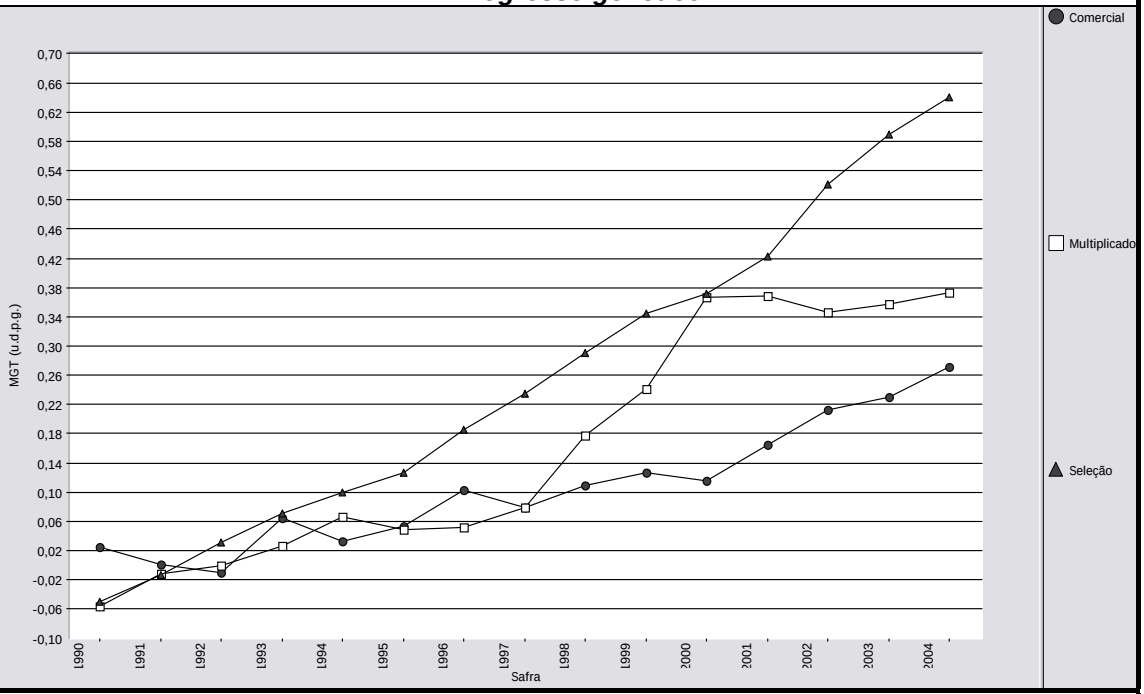
No período compreendido entre 1999 e 2003, o progresso genético foi de 0,25 unidades de desvio padrão genético (u.d.p.g.) para o rebanho *comercial*, 0,43 u.d.p.g. para o *multiplicador* e de 0,69 u.d.p.g. para o *seleção* (Figura 39). Estes resultados demonstram que, até o rebanho *comercial* vem sendo selecionado para crescimento, fertilidade e habilidade maternal. Os genes selecionados no topo da pirâmide (Figura 38) estão sendo efetivamente transferidos para o rebanho *comercial*.

Do total de 318.825 animais analisados, 75.596 (23,7%) foram oriundos de matrizes *multiplicadoras* e *comerciais*. 68.111 concepções foram oriundas de touros *seleção* nestas matrizes, representando 21,4% de todas as concepções e 90,1% das concepções destas matrizes (Figura 40). Estes resultados comprovam que ocorre, em larga escala, o fluxo gênico na pecuária de corte, para a Raça Nelore conforme indicado na Figura 38, devido à prática de se utilizar touros *seleção* em matrizes *multiplicadoras* e *comerciais*.

Concepções oriundas de touros *seleção* em matrizes *multiplicadoras* e *comerciais* vêm aumentando nos últimos anos (Figura 41), são 12,5%, 18,2% e 27,4% no primeiro, segundo e terceiro qüinqüênio, respectivamente (Divisão do número de animais, em casa qüinqüênio, do item *B* pelo *A* da Figura 41). A difusão da Inseminação Artificial (IA) foi a principal via de fluxo gênico, partindo dos rebanhos *seleção* até os *comerciais*, seguido da monta natural com uso de touros *seleção*. O uso de IA está em franca expansão.

Estes resultados podem ser explicados pelo grande número de touros jovens que o PMGRN – Nelore Brasil avaliou neste período e colocou no mercado. Geralmente touros com destaque nas DEPs são convidados pelas centrais de IA, o uso deste sêmen em rebanhos *comerciais* levam ao progresso genético do mesmo. Este fluxo tem potencial para aumentar nos próximos anos, dado que a oferta de sêmen de touros *seleção* cobre, apenas, 3% do rebanho *comercial* (PEREIRA, 2004).

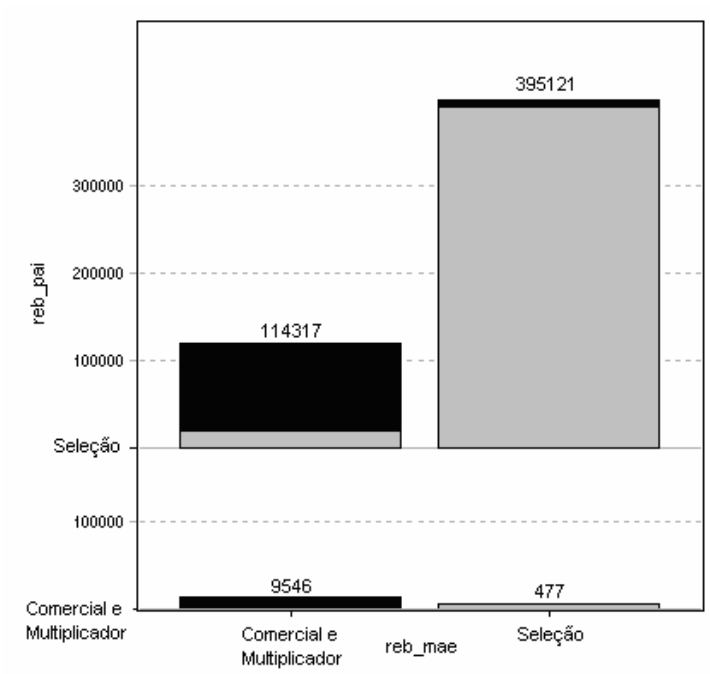
A. Progresso genético



B. Estatística descritiva

	N	$\bar{x}$	s	Máx.	Mín.
Comercial	113974	0,15	0,38	2,24	-1,91
Multiplicador	93032	0,28	0,41	2,32	-1,66
Seleção	427539	0,35	0,48	2,75	-1,89

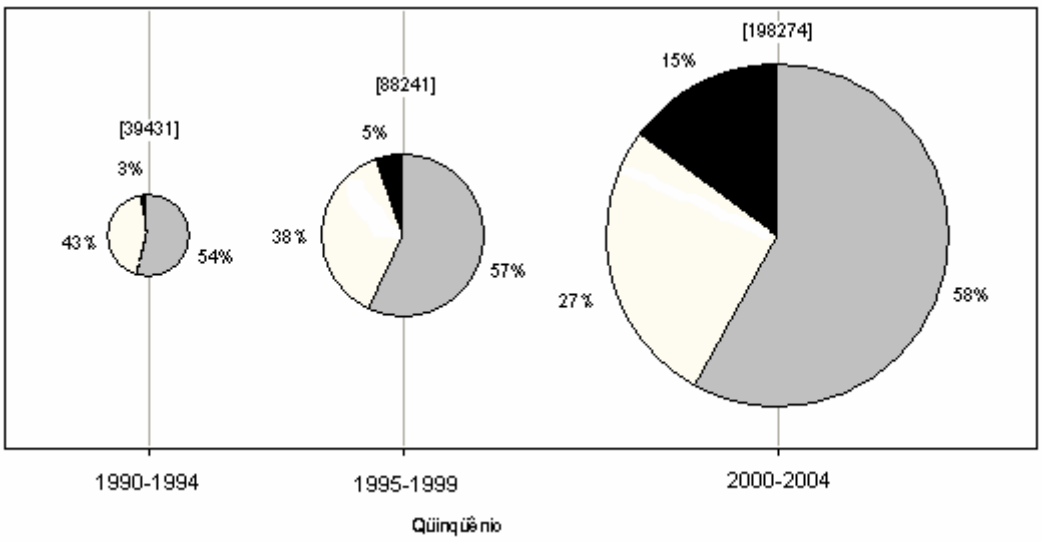
Figura 39 – A) Progresso genético para MGT por classe de categorias; B) Estatística descritiva.

A. Histograma	***B. Investigação de Detalhes***
	Barra ativa em Com. e Mult., Seleção: Contendo 114317 registros Informação de cor: Dividido conforme o 'Reb_Filho' Valor Número Percentagem Com. e Mult. 91181 84.9 Seleção 16159 15.1 Barra ativa em Seleção, Seleção: Contendo 395121 registros Informação de cor: Dividido conforme o 'Reb_Filho' Valor Número Percentagem Com. e Mult. 5863 1.5 Seleção 389258 98.5 Barra ativa em Com. e Mult., Com. e Mult.: Contendo 9545 registros Informação de cor: Dividido conforme o 'Reb_Filho' Valor Número Percentagem Com. e Mult. 9360 98.1 Seleção 186 1.9 Barra ativa em Seleção, Com. e Mult.: Contendo 477 registros Informação de cor: Dividido conforme o 'Reb_Filho' Valor Número Percentagem Com. e Mult. 201 42.1 Seleção 276 57.9

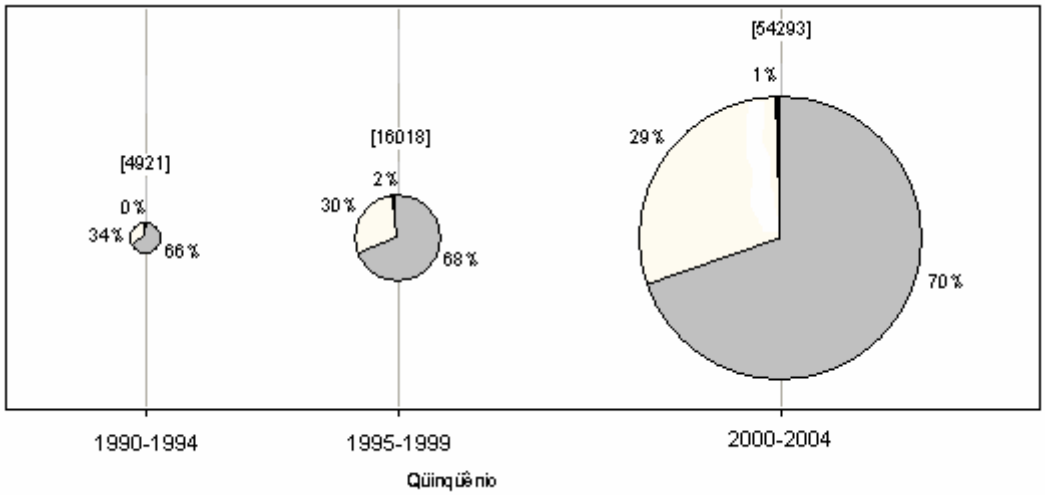
Legenda:
Cor por categoria da progênie: ··· *PO* e *POI* (*seleção*); ■ *CL* e *LA* (*comercial* e *multiplicador*);
Tamanho da barra por número de animais;
Rótulos: número de animais, porcentagem do mocho e porcentagem do padrão.

Figura 40 – A) Distribuição do número de animais concebidos pelos acasalamentos entre grupos de categorias dos progenitores; B) Investigação de Detalhes para categoria da progênie em cada grupo de acasalamentos.

A. Todo rebanho



B. Touros seleção em matrizes multiplicadoras e comerciais



Legenda:
Cor por tipo de acasalamento: ··· IA, ··· MN, ■ TE ou FIV;
Tamanho da pizza por número de animais;
Rótulos: número de animais e porcentagem do setor.

Figura 41 – Tipo de acasalamento por qüinqüênios.

5.3. Coeficiente de endogamia por categoria animal, safra e fazenda

O controle da endogamia foi diferente entre as categorias animais (Figura 42). Enquanto a média de endogamia para as categorias *CL* e *LA* não ultrapassa a 0,25% e *PO* não ultrapassa a 1,5%, a categoria *POI* passa de 3,5%. A endogamia dos animais *CL*, *LA* e *PO* praticamente não evoluiu na década analisada (safras 1994 a 2003), já os animais *POI* têm se tornado cada vez mais endogâmicos.

A elevada endogamia dos animais *POI* está relacionada ao fato de poucos touros e matrizes terem contribuído para formação do plantel desta categoria (Figura 43).

O controle da endogamia para animais na safra 2003 foi eficiente na maioria das fazendas analisadas (Figura 44). Dos 178 rebanhos analisados (restrição: mínimo de 100 animais/safra/fazenda com dados para F), apenas 4 fazendas (2,2%) apresentaram média de endogamia superior a 3%, 9 fazendas (5,1%) apresentaram média de endogamia entre 2% a 3% e 65 fazendas (36,5%) apresentaram média de endogamia entre 1 a 2%.

O grupo contendo as 13 fazendas com média para F ≥ 2% obteve progresso genético para MGT (0,61 u.d.p.g.) maior que o grupo contendo os 178 rebanhos (0,42 u.d.p.g.) (Figura 45). Esses resultados sugerem que a alta endogamia encontrada nessas 13 fazendas ainda não causou prejuízos no progresso genético ou que estes foram compensados pelos ganhos com a seleção. Todavia, essas fazendas têm que se preocupar com o controle da endogamia a médio e longo prazo, dado que, segundo Vozzi (2004), o crescente parentesco observado na Raça Nelore pode trazer conseqüências negativas no futuro para se atingir progressos genéticos em características de importância econômica. Com acasalamentos genéticos corretos é possível diminuir a média do coeficiente de endogamia destas fazendas.

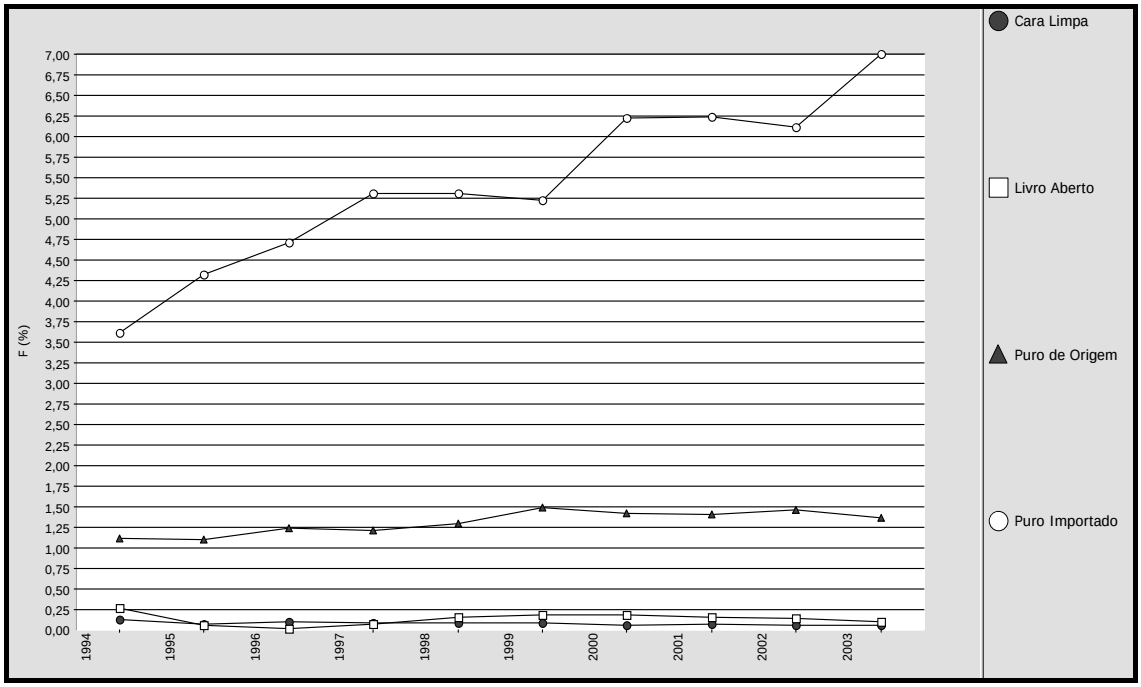


Figura 42 – Evolução da endogamia por categoria animal.

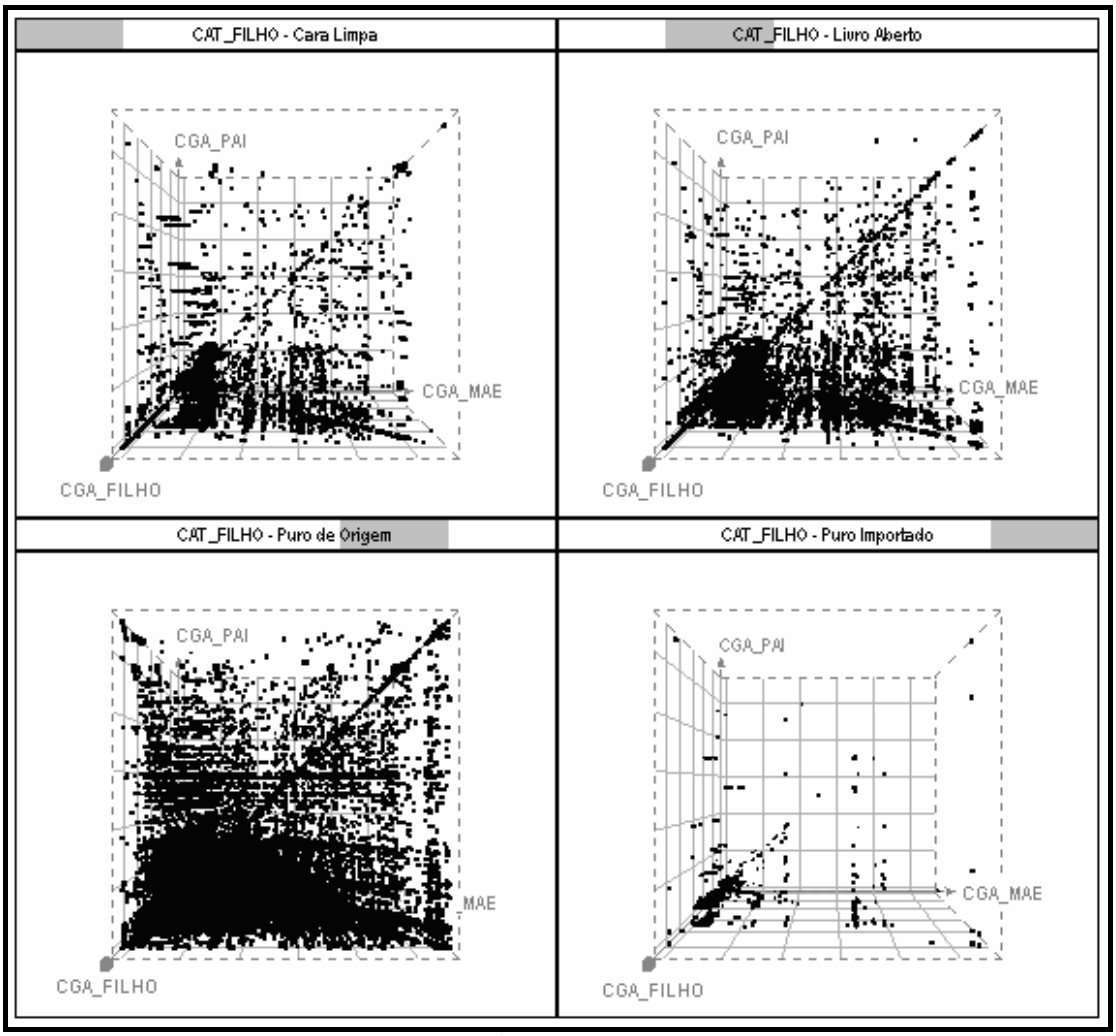


Figura 43 – Relação progênies e progenitores por categoria.

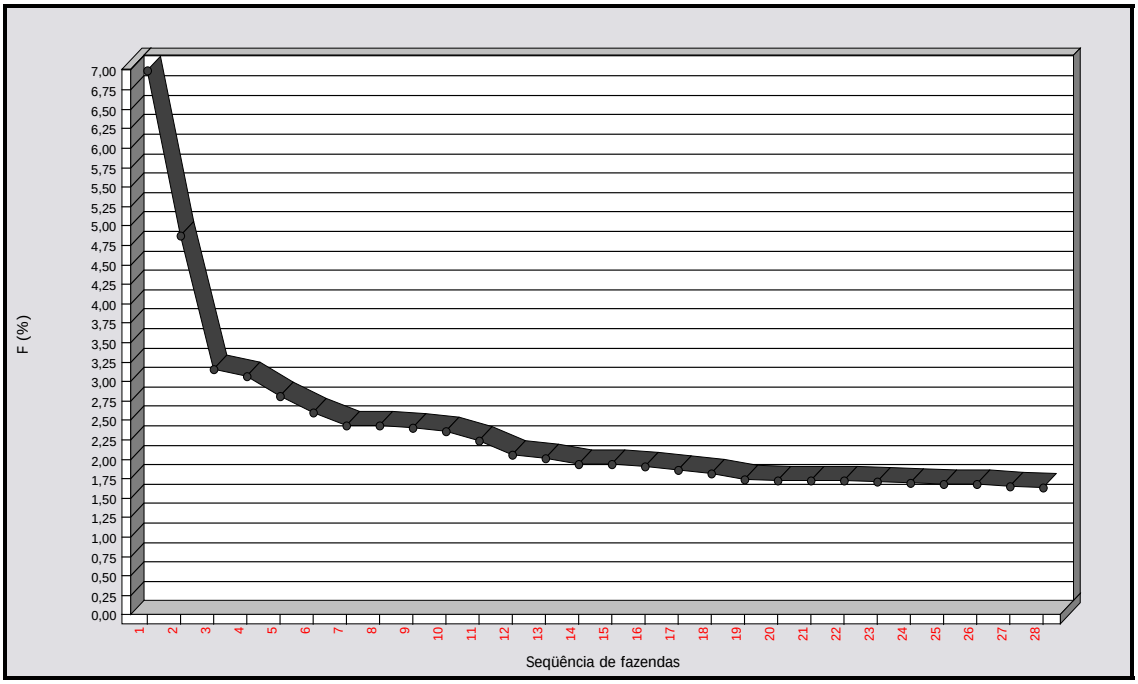
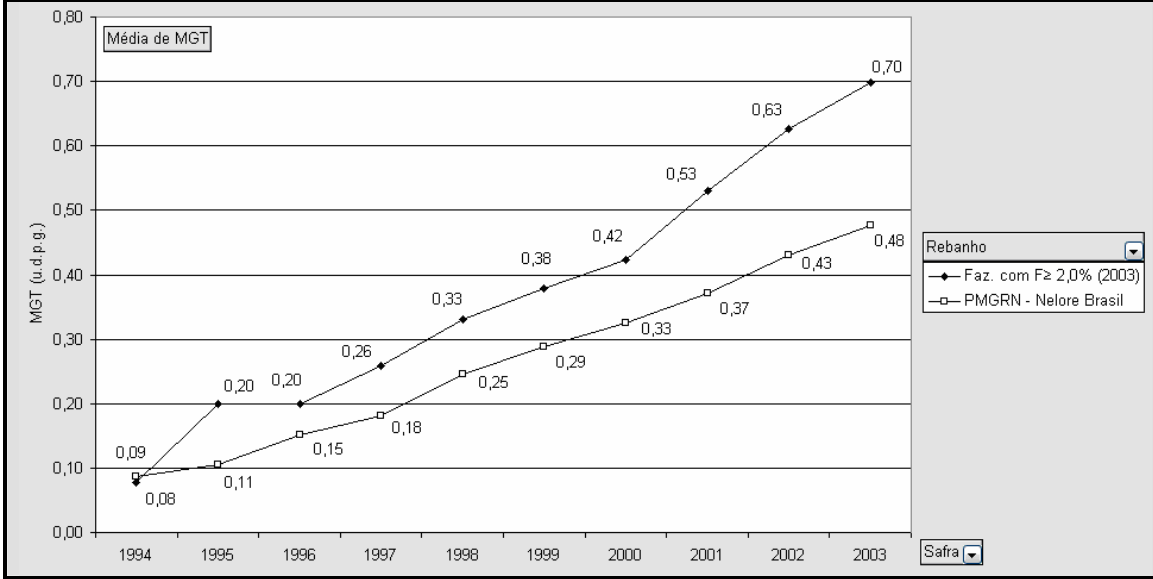


Figura 44 – Fazendas com maiores médias de endogamias para a safra 2003.
(Obs.: Os números não identificam fazendas).

A. Progresso Genético



B. Estatística descritiva

	Rebanho	N. fazendas	N. animais	$\bar{x}$	s	Máx.	Mín.
MGT	PMGRN - Nelore Brasil	178	71091	0,47	0,38	2,41	-1,66
	Faz. com F≥ 2,0% (2003)	13	3788	0,70	0,40	2,32	-1,39
F	PMGRN - Nelore Brasil	178	73709	0,86%	1,47%	31%	0%
	Faz. com F≥ 2,0% (2003)	13	3788	2,91%	2,79%	28%	0%

Restrição: Número mínimo animais/safra/fazenda com dados para F = 100.

Figura 45 – A) Progresso genético do MGT comparativo do PMGRN – Nelore Brasil com as fazendas que apresentaram média de F ≥ 2,0% na safra 2003; B) Estatística descritiva.

5.4. Identificação de padrões de seleção e acasalamento da Raça Nelore

Ao considerar todo o conjunto de dados, foram encontradas médias a altas correlações significativas (P<0,01), para cada par de DEPs, para as seis características analisadas (Tabela 10). Porém, o coeficiente de correlação de Pearson geralmente diminui ao restringir o conjunto de dados para animais TOP 25% (com exceção do coeficiente de correlação entre DIPP e DPE450). Porém, o mais interessante dentro de conjuntos de animais harmonicamente balanceados e de alto valor genético, não é apenas a correlação entre as DEPs, mas sim, a identificação do comportamento destes animais, daí a importância de se utilizar a mineração visual de dados.

Do total de 545.103 animais analisados, 10.846 (2,0%) são TOP 25% e pertencentes às safras 1994 e 2003. A distribuição destes animais por região é bastante irregular (Figura 46), mais de dois terços destes animais foram produzidos por fazendas nos estados de MS, SP, GO e MG. Porém, ao analisar a proporção dos animais TOP 25% pelo rebanho local, os estados de GO (7,4%) e RO (7,1%) destacam na produção de animais com este perfil.

A grande capacidade de propagação de material genético proporcionado pelas biotecnologias IA, TE e FIV (87% somando as três biotecnologias) foram responsáveis pela produção da maioria dos animais TOP 25% (Figura 47). Nos animais TOP 50%, a participação dessas biotecnologias são menores (59%, somando as três biotecnologias). Já, a maioria dos animais BOTTON 50% foram produzidos à partir de MN (65%). Esses resultados demonstram o bom gerenciamento de biotecnologias reprodutivas por parte do PMGRN – Nelore Brasil.

Os rebanhos *seleção*, representados por animais *PO* (79%), foram os maiores responsáveis pela produção destes animais TOP 25% (Figura 48). Houve notável participação da categoria *LA* (14%), possivelmente pelo uso de sêmen de touros *PO* nestas matrizes (Figura 40).

Ao relacionar a identificação dos animais TOP 25% com as fazendas produtoras e potencial genético (Figura 49), os padrões visuais revelam que algumas fazendas têm maior habilidade em produzir animais com este perfil, refletindo em alto MGT dos mesmos. Embora a maioria das fazendas se dedique a rebanhos *PO*, várias criatórios conseguem produzir animais *CL* e *LA* com este perfil. O equilíbrio entre sexos está presente na maioria das fazendas.

O perfil genético dos animais TOP 25% (Figura 50) demonstra que neste grupo, os animais expressam alto MGT (à exceção de uma única matriz *CL*, todos os outros animais são, pelo menos, TOP 25% para MGT). Esses resultados demonstram que é possível incluir, como critérios de seleção de um rebanho, animais harmônicos para características reprodutivas, além da habilidade materna, crescimento e fertilidade (contempladas pelo MGT). Apenas 109 dos animais (1,01%) TOP 25% possuem F ≥ 10,0% e 18 animais (0,17%) possuem F ≥ 20%, indicando que há possibilidade de seleção para animais de alto e balanceado valor genético sem comprometer o rebanho com elevada consangüinidade.

Grande parte dos animais TOP 25% parecem não terem sido aproveitados para reprodução, dentro do PMGRN – Nelore Brasil (Figura 51, Tabela 11 e Tabela 12), dado que 5.198 touros (97,9% dos TOP 25%) e 4.152 matrizes (76,9% das TOP 25%) não deixaram descendentes avaliados no rebanho. Eles deveriam ser melhores aproveitados pelos criadores do programa. Porém, a última situação observada da maioria destes animais está declarada como “ativo” ou “vendido”, a minoria como “morto” ou “descarte”, ou seja, não foram descartados do processo reprodutivo e sim, utilizados como reprodutores em rebanhos de clientes dos criadores participantes do PMGRN – Nelore Brasil (vários criadores esquecem de declarar em suas bases de dados, a venda dos animais). Portanto, analisando sobre outra ótica, esses animais são fontes de receitas das fazendas, fornecendo os reprodutores que a cadeia produtiva da pecuária de corte necessita para ser eficiente.

Observando nas Tabelas 11 e 12, a percentagem de animais por classe (colunas) para cada intervalo do número de progênies avaliadas (linhas), fica evidente que animais de alto e balanceado valor genético tem maior probabilidade de se destacarem como reprodutores. Esta é uma vantagem comercial para as fazendas que produzem este perfil de animal, pela maior probabilidade de venda de sêmen e embriões.

Este conjunto de resultados demonstra o poder que a mineração visual de dados tem em produzir informações e conhecimentos para assessorar criadores, de um programa de melhoramento genético, na busca de maior eficiência nos processos de seleção e acasalamento dos animais.

Tabela 10 – Coeficiente de correlação de Pearson para os conjuntos de dados: todos os animais (linha superior) e TOP 25% (linha inferior).

	MP120	***DP120***	***DP450***	***DPE450***	***DIPP***	***DPAC***
MP120	1,00000 1,00000	0,28253** 0,17914**	0,36126** 0,23978**	0,15727** -0,09293**	-0,20262** -0,05556**	0,47713** 0,17180**
DP120		1,00000 1,00000	0,80538** 0,52587**	0,37458** 0,17668**	-0,21843** -0,05757**	0,28701** -0,16101**
DP450			1,00000 1,00000	0,41057** 0,17598**	-0,30890** -0,10418**	0,39435** 0,08410**
DPE450				1,00000 1,00000	-0,18341** -0,34939**	0,23184** 0,09824**
DIPP					1,00000 1,00000	-0,58666** -0,39480**
DPAC						1,00000 1,00000

* Significativo a 5%;
** Significativo a 1%.

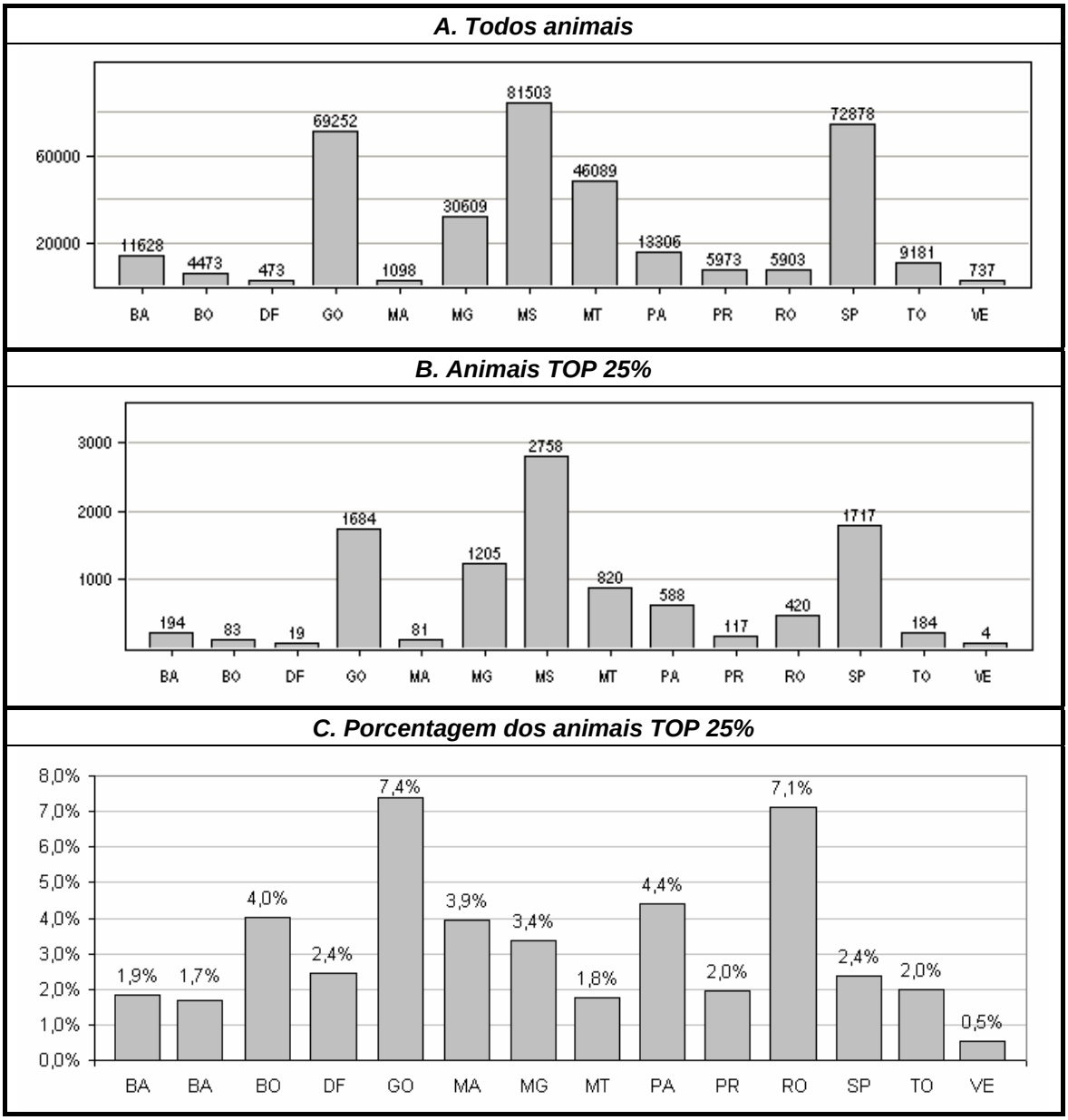


Figura 46 – Distribuição dos animais nascidos entre 1994 e 2003 por estado brasileiro e outros países: A) Distribuição de todos animais; B) Distribuição dos animais TOP 25%; C) Porcentagem dos animais TOP 25%.

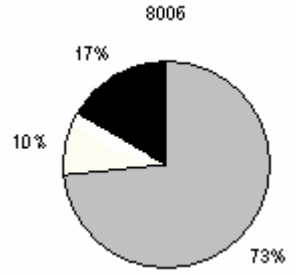
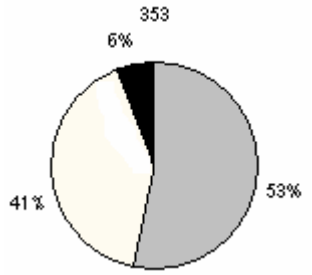
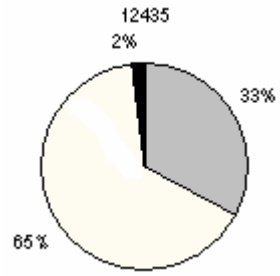
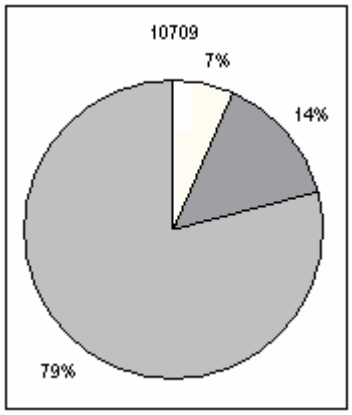
A. TOP 25%	*B. TOP 50%*	*C. BOTTON 50%*
		
Legenda: Cor por tipo de acasalamento: ··· IA, ··· MN, ■ TE ou FIV; Rótulos: número de animais e porcentagem do setor.		

Figura 47 – Distribuição dos animais por tipo de acasalamento e classe.



Legenda:

Cor por categoria:
··· *CL*,
··· *LA*,
··· *PO*,

Rótulos: número de animais e porcentagem do setor.

Figura 48 – Distribuição dos animais TOP 25% por categoria.

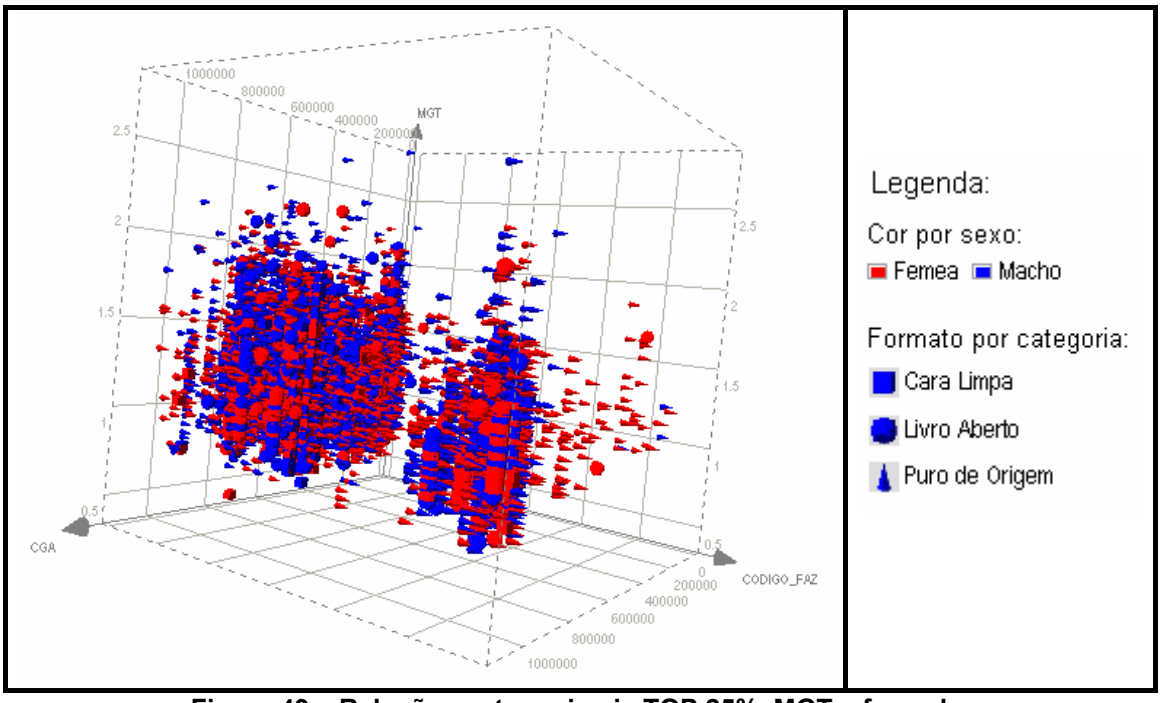


Figura 49 – Relações entre animais TOP 25%, MGT e fazenda.

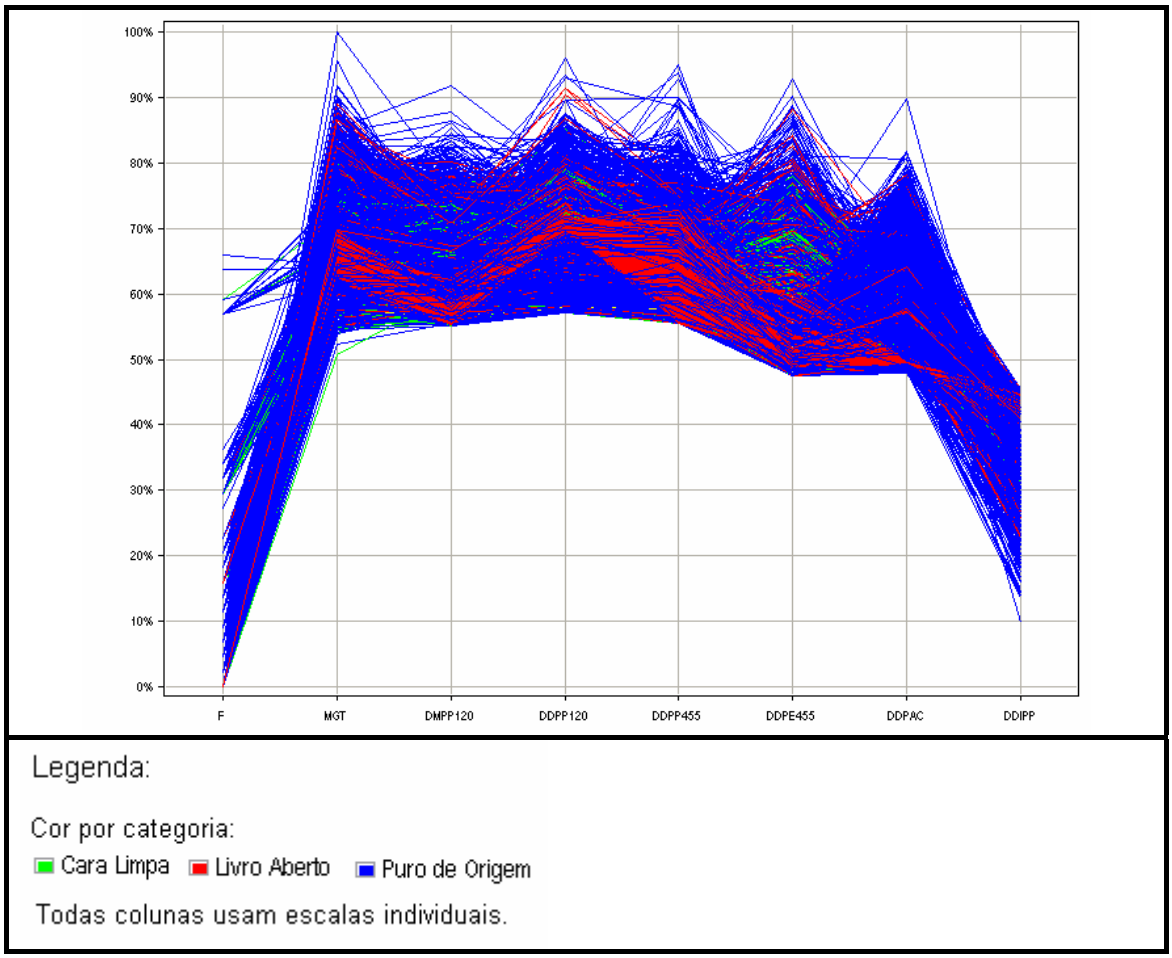


Figura 50 – Perfil genético dos animais TOP 25%.

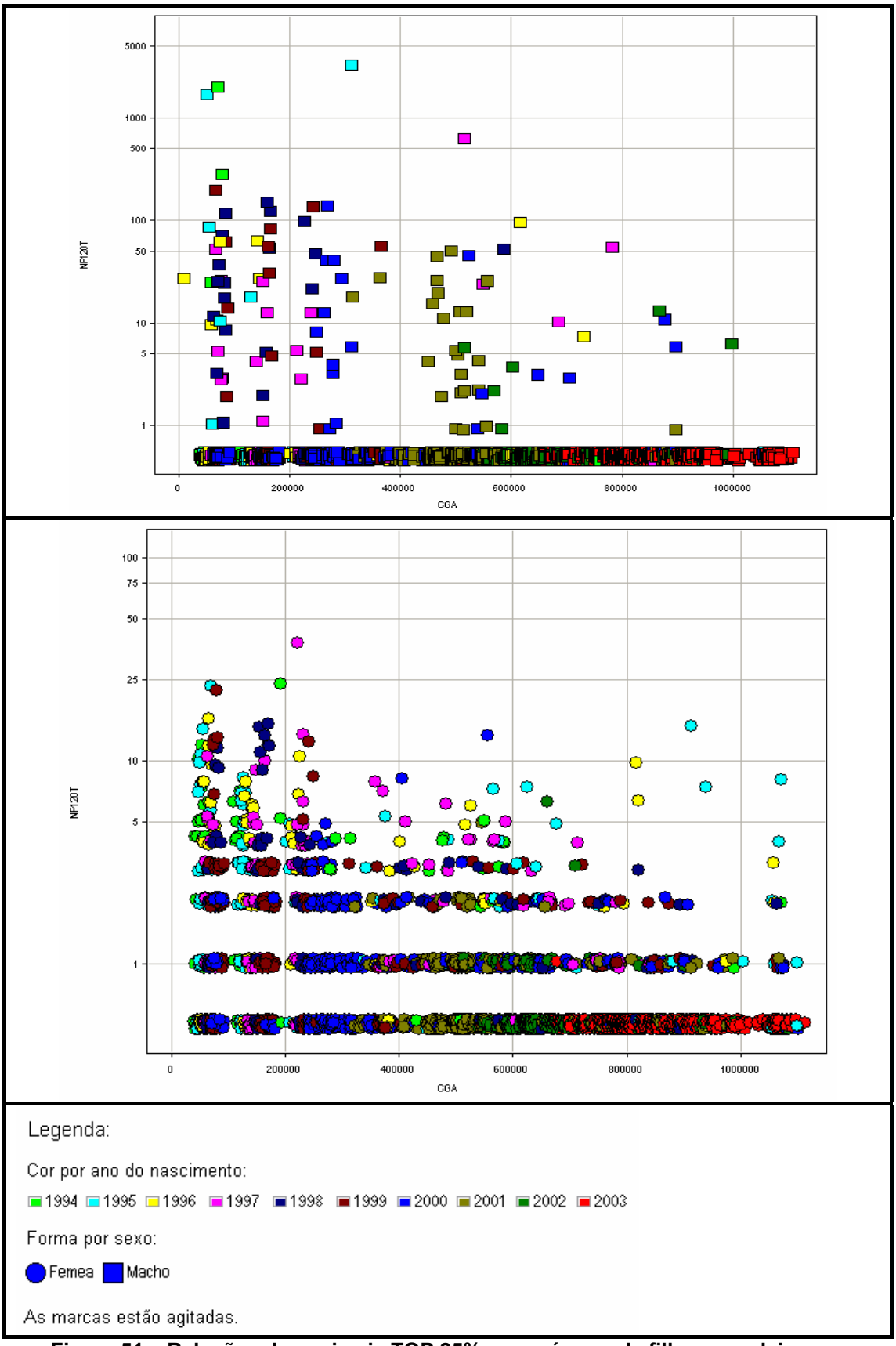


Figura 51 – Relações dos animais TOP 25% com número de filhos que deixaram avaliados no rebanho (NF120).

Tabela 11 – Relação entre número de filhos avaliados e classe dos touros.

NF120	*TOP 25%*		*TOP 25% e/ou TOP 50%*		*TOP 50%*		*BOTTON 50%*		*Todos*
Intervalo	*N*	*P*	*N*	*P*	*N*	*P*	*N*	*P*	*N*
0	5.198	2,92%	25.106	14,11%	247	0,14%	10.014	5,63%	177.921
[1;9[	49	5,62%	182	20,87%	1	0,11%	26	2,98%	872
[10;99[	52	4,36%	207	17,35%	0	0,00%	7	0,59%	1.193
[100;999[	9	4,29%	40	19,05%	0	0,00%	1	0,48%	210
[1000;10000[	3	17,65%	5	29,41%	0	0,00%	0	0,00%	17
Total	5.311		25.540		248		10.048		180.213
Total NF120(1)	113		434		1		34		2.292

N: número de touros; *P*: percentagem em relação à coluna *Todos*; (1) Número de animais com NF120 ≥ 1.

Tabela 12 – Relação entre número de filhos avaliados e classe das matrizes.

NF120	*TOP 25%*		*TOP 25% e/ou TOP 50%*		*TOP 50%*		*BOTTON 50%*		*Todos*
Intervalo	*N*	*P*	*N*	*P*	*N*	*P*	*N*	*P*	*N*
0	4.152	2,83%	21.011	14,34%	202	0,14%	9.402	6,42%	146.528
[1;9[	1.218	2,06%	7.125	12,03%	39	0,07%	4.897	8,26%	59.251
[10;99[	30	8,52%	97	27,56%	0	0,00%	4	1,14%	352
Total	5.400		28.233		241		14.303		206.131
Total NF120(1)	1.248		7.222		39		4.901		59.603

N: número de matrizes; *P*: percentagem em relação à coluna *Todos*; (1) Número de animais com NF120 ≥ 1.

Inúmeras fazendas do PMGRN – Nelore Brasil utilizaram da prática de TE e FIV. Para as safras compreendidas entre 1994 a 2003, 16.384 animais foram concebidos (3,5 % do total) por estas biotecnologias reprodutivas. Segundo Elias (2003), com o uso de biotecnologias reprodutivas como TE e FIV, uma matriz passa a ter o potencial de produzir, em média, uma cria por mês e uma por semana, respectivamente. Portanto é necessário que o PMGRN – Nelore Brasil monitore o uso dessas biotecnologias e tome ações quando detectar eventuais falhas.

De todas as matrizes nascidas a partir de 1990, aproximadamente 1500 delas foram destinadas a TE ou FIV, no período de 1994 a 2003 e, apenas, 12 delas se enquadram como matrizes BOTTON 50% (Figura 52). Outro padrão interessante que pode ser observado é o mês de acasalamento, que ocorreu preferencialmente em janeiro, ou seja, dentro da estação de acasalamento utilizada na maioria das fazendas participantes do PMGRN – Nelore Brasil, propiciando a formação dos lotes de manejo. Dessas matrizes, 4 delas foram oriundas da TE ou FIV, indicando o uso dessas biotecnologias reprodutivas nessas fêmeas pelo simples fato de terem sido concebidas pela mesma prática, fato observado em animais destinados a pistas de julgamento.

Essas 12 matrizes BOTTON 50% contribuíram com 134 progênies avaliadas por meio de TE ou FIV (0,82 %) e mais 33 (0,20%) progênies por meio de outros tipos de acasalamentos (Figura 53). O primeiro relato dessa biotecnologia reprodutiva, nas matrizes BOTTON 50% data de 1997, coincidindo com o início da FIV do Brasil, sendo que a casuística aumentou a partir de 2001.

As 12 matrizes BOTTON 50% pertencem a 9 fazendas, porém as 167 progênies foram concebidas em 23 diferentes rebanhos (Figura 54). Esses números demonstram que algumas fazendas podem ter errado na aquisição destes embriões. Cabe ressaltar a importância da seleção objetiva de fêmeas para TE ou FIV, visto a velocidade de propagação genética entre diferentes rebanhos com o uso dessas biotecnologias.

A maioria das progênies dessas matrizes *críticas* apresentam um perfil genético com baixas DEPs para todas as características analisadas (Figura 55). As medianas das DEPs dessas progênies comprovam seu baixo valor genético (Tabela 13).

Com a popularização das biotecnologias reprodutivas, TE e FIV, algumas matrizes BOTTON 50% foram submetidas a esta prática. Realmente, aproveitar progenitoras de baixo potencial genético pode acarretar no nascimento maciço de animais indesejáveis. Os padrões aqui extraídos, por um aplicativo de mineração visual de dados, podem ajudar as fazendas a entenderem e evitarem eventuais erros de manejo reprodutivo. A incorporação destes conhecimentos, junto aos criadores, podem auxiliá-los no uso das biotecnologias reprodutivas de TE e FIV.

Essa tecnologia permite, ainda, auxiliar o criador a diminuir o risco de seu investimento, na aquisição de embriões, pela análise do perfil genético dos progenitores e progênies, que estes já deixaram avaliados no rebanho.

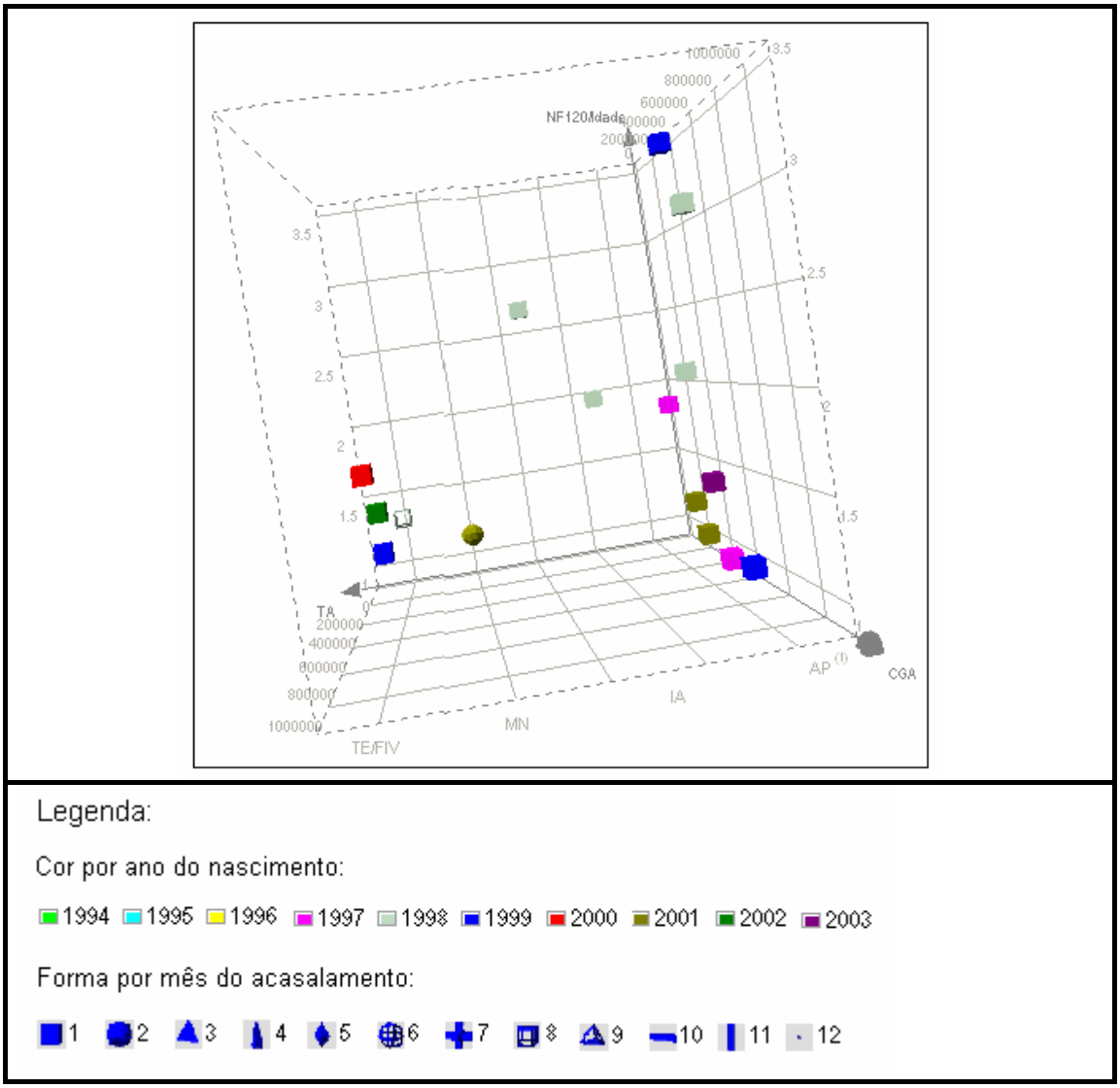


Figura 52 – Relações das matrizes BOTTON 50% que foram submetidas a TE ou FIV com número de filhos avaliados aos 120 dias de idade (NF120) e tipo de acasalamento (TA) que deram origem a essas matrizes.

(1) AP indica animais avaliados pela matriz de parentesco, portanto o TA é desconhecido.

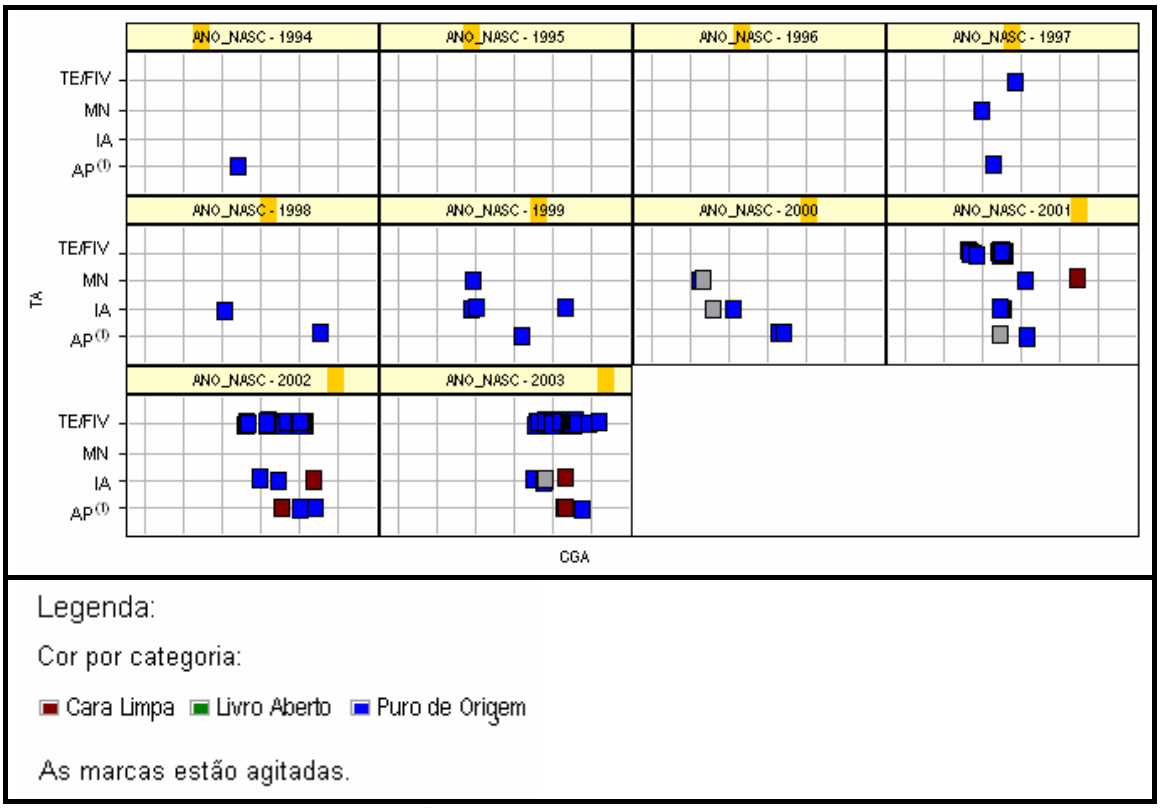


Figura 53 – Relações das progênies oriundas das matrizes BOTTON 50% com o tipo de acasalamento que originaram as progênies, por safra.

(1) AP indica animais avaliados pela matriz de parentesco, portanto o TA é desconhecido.

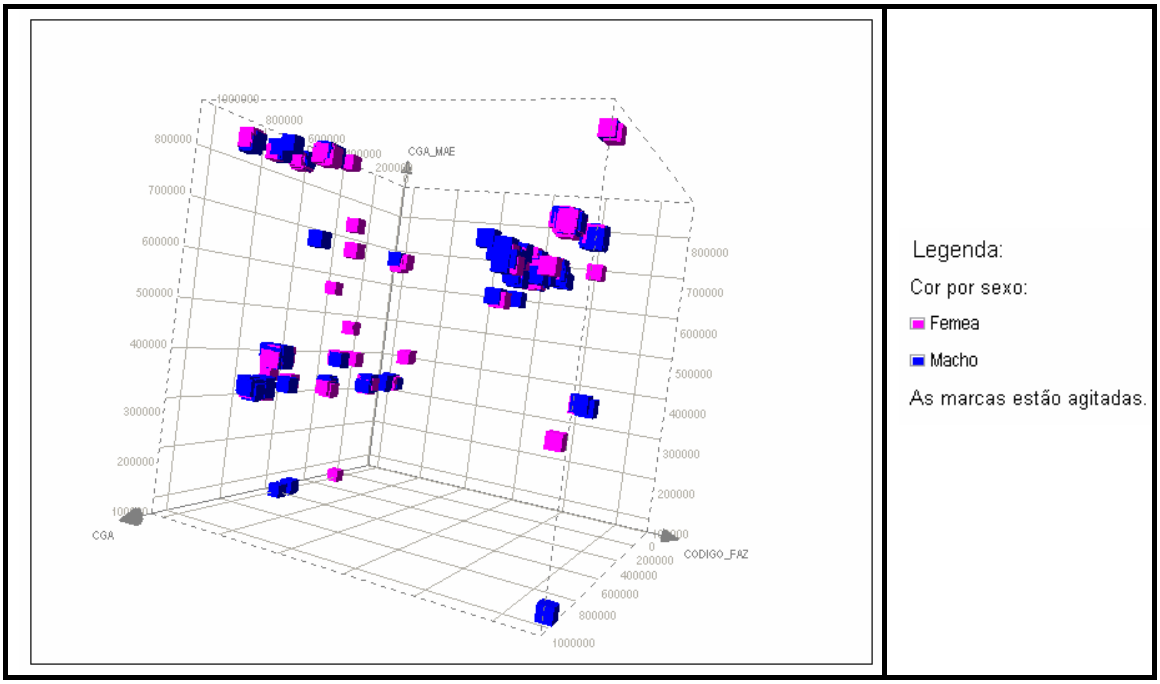


Figura 54 – Relações das progênies oriundas das matrizes BOTTON 50% que foram submetidas a TE ou FIV com a matriz e a fazenda onde ocorreu o parto.

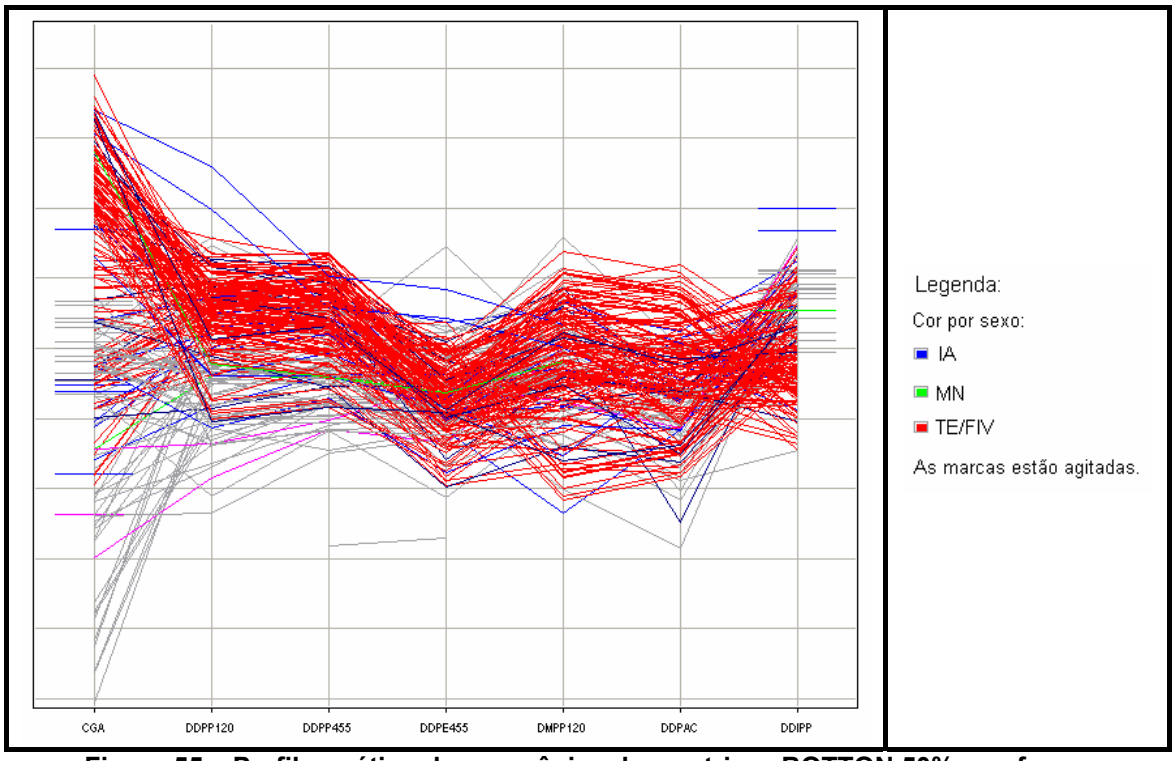
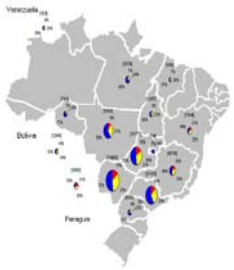
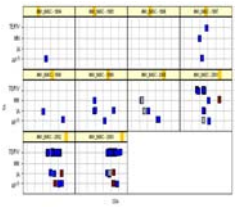
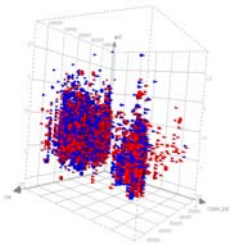
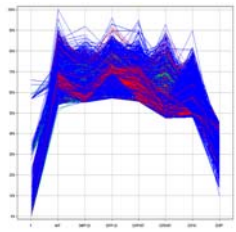
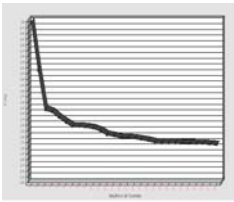


Figura 55 – Perfil genético das progênies das matrizes BOTTON 50% que foram submetidas a TE ou FIV.

Tabela 13 – Valores máximos, mínimos e medianas das DEPs de animais provenientes de matrizes BOTTON 50% que foram submetidas a TE ou FIV.

DEP	Máximo		Mediana		Mínimo	
	Valor	Percentil(2)	Valor	Percentil(2)	Valor	Percentil(2)
DP120	5,04	2,0%	1,45	40,0%	-4,38	-50,0%
DP450	9,95	10,0%	4,96	40,0%	-14,49	-50,0%
DPE450	1,02	1,0%	-0,13	-50,0%	-0,87	-50,0%
MP120	2,06	10,0%	0,16	-50,0%	-2,73	-50,0%
DPAC	4,77	5,0%	2,00	40,0%	-5,04	-50,0%
DIPP(1)	-0,66	10,0%	-0,19	50,0%	0,38	-50,0%

(1) Valores negativos são esperados;
(2) Valores positivos indicam animais TOP e negativos, BOTTON;

CONCLUSÕES E PROPOSTOS FUTURAS

6. CONCLUSÕES E PROPOSTOS FUTURAS

6.1. Conclusões

O uso da tecnologia da informação, de consulta *OLAP* e de mineração visual de dados, apresenta eficiência e eficácia na caracterização da estrutura populacional da Raça Nelore. Os padrões extraídos indicam nichos de mercado para os criadores do programa de melhoramento genético, ou seja, demonstram as preferências do mercado consumidor dos animais avaliados.

As tecnologias utilizadas transformam impressões em fatos. Realmente, há o fluxo gênico do rebanho *seleção* para o *multiplicador* e o *comercial*, principalmente pelo uso de touros *PO* em matrizes *LA* e *CL*. A crescente participação da IA levou a rápida difusão do material genético dos touros *PO*. O fluxo gênico foi responsável pelo progresso genético do rebanho *comercial*. Este fluxo gênico incrementa o progresso genético do rebanho destinado à produção de carne e aumenta a competitividade da pecuária de corte.

Embora o controle da endogamia na Raça Nelore seja preocupante devido à presença de poucos fundadores da raça, houve bom controle da endogamia nas categorias *CL*, *LA* e *PO*, na maioria das fazendas.

O PMGRN – Nelore Brasil foi eficiente em levar informações e assessorar seus criadores na seleção e acasalamento dos animais. A mineração visual de dados tem potencial para incrementar o processo seletivo e maximizar o progresso genético do rebanho.

A maioria das fazendas participantes do PMGRN – Nelore Brasil foram eficientes na escolha de matrizes destinadas às biotecnologias de TE e FIV. Investigação de perfil genético e reprodutivo, proporcionada pela mineração visual de dados, é uma forma de análise de risco de investimento para aquisição de embriões.

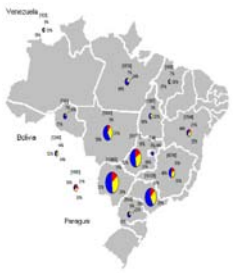
O *Nelore Business Intelligence* é eficiente como ferramenta gerencial do PMGRN – Nelore Brasil e fonte de dados para pesquisa científica. Ele proporciona um ambiente eficaz para extração de informação e conhecimento a respeito da avaliação genética do PMGRN – Nelore Brasil.

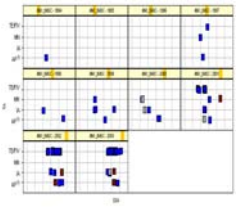
6.2. Propostas futuras

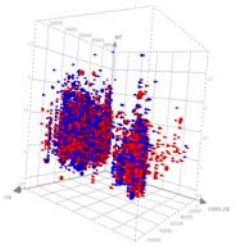
Dada a importância do *Data Warehouse* para o PMGRN – Nelore Brasil, o *Nelore Business Intelligence* deve sofrer evoluções constantemente, duas ações imediatas são:

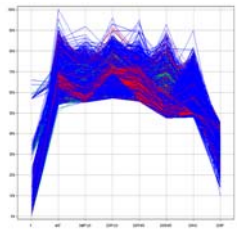
- **Disponibilização *on-line* do *Data Warehouse* (*Data Webhouse*):** Disponibilizar pela *internet* ou *intranet* para Processamento Analítico *On-Line* (*OLAP*), aos usuários do PMGRN – Nelore Brasil (pesquisadores e técnicos). A transformação do *Nelore Business Intelligence* num *Data Webhouse* será capaz de levar informações úteis e necessária ao gerenciamento dos rebanhos participantes. Também é necessário, o treinamento da comunidade de usuários, para uso de aplicativos *OLAP*;
- **Criação de novos *Data Marts*:** Para melhorar as condições de gerenciamento do PMGRN – Nelore Brasil, outros *Data Marts* devem ser implementados, como controle financeiro e relatório de consultorias às fazendas. Novos *Data Marts* serão valiosos repositórios de dados para pesquisas futuras.

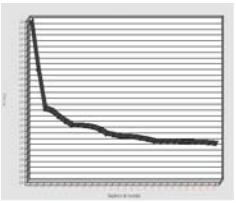
O *Nelore Business Intelligence* deve, ainda, ser utilizado para pesquisa com outros níveis de mineração de dados, como extração automática de padrões e construção de modelos, gerando novos conhecimentos úteis e válidos.











REFERÊNCIAS

REFERÊNCIAS

AHLBERG, C. Spotfire: an information exploration environment. **ACM SIGMOD Records**, New York, v. 25, n. 4, p. 25-29, 1996.

AHLBERG, C.; SHNEIDERMAN, B. Visual Information Seeking: tight coupling of dynamic query filters with starfield displays. In: PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS: CELEBRATING INTERDEPENDENCE, S., 1994, Boston. **Anais...** New York: ACM Press, 1994, p. 313-317. Disponível em: http://portal.acm.org/. Acesso em: 10 fev. 2006.

ALBUQUERQUE, L.G.; BERGMANN, J.A.G; OLIVEIRA, H.N.; TONHATI, H.; LÔBO, R.B. Princípios de Avaliação Genética. In: WORKSHOP SELEÇÃO EM BOVINOS DE CORTE, S., 2003, Salvador. **Anais...** Ribeirão Preto: ANCP, 2003. CD-ROM.

ALMEIDA, M.O.; MENDONÇA NETO, M.G. Uso de Interfaces Abundantes em Informação para Mineração Visual de Dados. (Relatórios técnicos: RT-NUPERC-2001-5/p). Bahia: Universidade Salvador, 2001.

AMARAL, F.C.N. **Data Mining: técnicas e aplicações para o marketing direto**. São Paulo: Berkeley, 2001.

ASSOCIAÇÃO BRASILEIRA DOS CRIADORES DE ZEBU [ABCZ]. O Zebu Nunca Parou de Crescer. Uberaba. Disponível em: http://www.abcz.org.br/. Acesso em: 11 fev. 2006.

ASSOCIAÇÃO NACIONAL DE CRIADORES E PESQUISADORES [ANCP]. Sumários. Ribeirão Preto. Disponível em: http://www.ancp.org.br/. Acesso em 14 fev. 2006.

BARIONI, M.C.N. Visualização de Operações de Junção em Sistemas de Base de Dados para Mineração de Dados 64 p.; 2002. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2002.

BATISTA, G.E.A.P. Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado 204 p.; 2003. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003.

BERGMANN, J.A.G. Objetivos e Critérios de Seleção. In: WORKSHOP SELEÇÃO EM BOVINOS DE CORTE, S., 2003, Salvador. **Anais...** Ribeirão Preto: ANCP, 2003. CD-ROM.

BOLDMAN, K.G.; VAN VLECK, L.D.; KRIESE, L.M.; KACHMAN, S. **MTDFREML – User's Guide**. Washington: USDA-ARS, 1995.

BREIMAN, L. Bagging Predictors. **Machine Learning**, New York, v. 24, p. 123-140, 1996.

CARNEIRO, P.S.C.; TORRES, R.A.; EUCLYDES, R.F.; SILVA, M.A.; LOPES, P.S.; CARNEIRO, P.L.S.; TOORES FILHO, R. A. T. Efeito da Conexidade de Dados sobre a Acurácia dos Testes de Progênie e Performance. **R. Bras. Zootec.**, Viçosa, v. 30, n. 2, 2001.

CENTRO DE ESTUDOS AVANÇCADOS EM ECONOMIA APLICADA [CEPEA]. PIB no Agronegócio Brasileiro. Piracicaba. Disponível em: http://www.cepea.esalq.usp.br/. Acesso em: 31 dez. 2005.

CHEN, M.S.; HAN, J.; YU, P.S. Data Mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering**, Los Alamitos, v. 8, n. 6, p. 866-883, 1996.

COLLIATE, G. OLAP, Relational and Multidimensional Database Systems. **ACM SIGMOD Records**, New York, v. 25, n. 3, 1996.

COME, G. Contribuição ao Estudo da Implementação de Data Warehousing: um caso no setor das telecomunicações 133 p.; 2001. Dissertação (Mestrado) – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2001.

COREY, M.J.; ABBEY, M.S.; ABRAMSON, I.; TAUB, B. **Oracle 8i Data Warehouse**. Rio de Janeiro: Campus, 2001.

CRAVEN, M.W.; SHAVILIK, J.S. Extracting Comprehensible Concept Representation from Trained Neural Networks. In: PRESENTED AT THE IJCAI WORKSHOP ON COMPREHENSIBILITY IN MACHINE LEARNING, S., 1995, Montreal. **Anais...** Madison: University of Wisconsin, 1995, p. 1-15. Disponível em: http://www.cs.wisc.edu/. Acesso em: 15 jan. 2006.

CRUZ, C.D.; REGAZZI, A.J. **Modelos Biométricos Aplicados ao Melhoramento Genético**. 2ª ed. Viçosa: UFV, 1997.

DISCOVERER. Oracle Discoverer Administration Edition: administration guide. (Release 4.1 for Windows). Redwood Shores: Oracle Corporation, 2000a.

DISCOVERER. Oracle Discoverer Plus: user's guide. (Release 4.1 for Windows). Redwood Shores: Oracle Corporation, 2000b.

EICK, S.G. Visual Discovery and Analysis. **IEEE Transactions on Visualization and Computer Graphics**, Los Alamitos, v. 6, n. 1, p. 44-58, 2000.

ELIAS, F.P. Avaliação da Fertilidade In Vitro de Touros do PMGRN – USP e dos Estágios de Desenvolvimento dos Embriões Produzidos 48 p.; 2003. Monografia (Trabalho de Conclusão) – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2003.

EXCEL. Microsoft Office Excel. São Paulo: Microsoft Informática Ltda., 2003.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to Quantitative Genetics**. Edinburgh: Longman, 1994.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH. From Data Mining to Knowledge Discovery. **AI Magazine**, Menlo Park, v. 17, p. 37-54, 1996.

FAYYAD, U.M.; UTHURUSAMI, R. Evolving Data Mining Into Solutions for Insights. **Communications of ACM**, New York, v. 45, n. 8, p. 28-31, 2002.

FORTES, G. Sumários Avançam em Consistência. **DBO Genética**, São Paulo, p. 24-25, 2005.

FOULLEY, J.L.; BOUIX, J.; GOFFINET, B.; ELSEN, J.M. Connectedness in Genetic Evaluation. In: GIANOLA, D.; HAMMOND, K. **Advances in Statistical Methods for Genetics Evaluations**. Berlin: Speinger-Verlad, 1990.

GANESH, M.; HAN E-H.; KUMAR, V.; SHEKHAR, S.; SRIVASTAVA, J. Visual Data Mining: framework and algorithm development. (Technical Reports TR-96-021). Minneapolis: University of Minnesota, 1996.

GARDNER, S.R. Building the Data Warehouse. **Communications of the ACM**, New York, v. 41, n. 9, p. 52-60, 1998.

GATZIU, S.; VAVOURAS, A. Data Warehousing: concepts and mechanisms. **Informatik – Informatique**, Zurich, v. 1, p. 8-11, 1999.

GOEBEL, M.; GRUENWALD, L. A Survey of Data Mining and Knowledge Discovery Software Tolls. **ACM SIGKDD Explorations**, New York, v. 1, n. 1, p. 20-33, 1999.

GOLDEN, B.L.; SNELLING, W.M.; MALLINCKRODT, C.H. Animal Breeder's Toll Kit User's Guide and Reference Manual TK3/TKBLUP. (Tech. Bulletin LTB92-2). Fort Collins: Colorado State University, 1995.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro: Campus, 2005.

GOLFARELLI, M.D.; MAIO, D.; RIZZI, S. Conceptual Design of Data Warehouses from E/R Schemes. In: PROCEEDINGS OF THE THIRTY-FIRST ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, S., 1998, Kona. **Anais...** Washington: IEEE Computer Society, 1998. Disponível em: http://portal.acm.org/. Acesso em: 13 jan. 2006.

GUIZZO, É. Bit Bem Passado. **Exame**, São Paulo, v. 74, 2001.

INMON, W.H. **Como construir o Data Warehouse**. Rio de Janeiro: Campus, 1997.

INMON, W.H.; TERDEMAN, R.H.; IMHOFF, C. **Data Warehousing: como transformar informações em oportunidade de negócios**. São Pulo: Berkeley, 2001.

INSELBERG, A. Multidimensional Detective. In: PROCEEDINGS OF THE 1997 IEEE SYMPOSIUM ON INFORMATION VISUALIZATION, S., 1997, Raanana. **Anais…** Washington: IEEE Computer Society, 1997. p. 100-107. Disponível em: http://portal.acm.org/. Acesso em 11 fev. 2006.

INSELBERG, A.; DIMSDALE, B. Parallel Coordinates: a toll for visualizing multi-dimensional geometry. In: PROCEEDINGS OF THE 1ST CONFERENCE ON VISUALIZATION '90, S., 1990, San Francisco. **Anais…** Los Alamitos: IEEE Computer Society Press, 1990. p. 361-378. Disponível em: http://portal.acm.org/. Acesso em: 11 fev. 2006.

INSTITUTO DE ECONOMIA APLICADA [IEA]. Comércio Exterior. São Paulo. Disponível em: http://www.iea.sp.gov.br/. Acesso em: 09 fev. 2006.

KEIM, D.A. Designing Pixel Oriented Visualization Techniques: theory and applications. **IEEE Transactions on Visualization and Computer Graphics**, Los Alamitos, v. 6, n. 1, p. 59-78, 2000.

KEIM, D.A. Visual Exploration of Large Data Sets. **Comunications of the ACM**, New York, v. 44, n. 8, p. 38-44, 2001.

KEIM, D.A., KRIEGEL, H.P. Visualizations Techniques for Mining Large Databases: a comparison. **IEEE Transactions on Knowledge and Data Engineering**, Los Alamitos, v. 8, n. 6, p. 923-938, 1996.

KNOWLES, J. Explore Data Warehousing. **Datamation**, Cambridge, v. 42, n. 18, p. 30, 1996.

LIMA, F.P. Expansão da Raça Nelore no Brasil. In: LÔBO, R.B.; BEZERRA, L.A.F.; OLIVEIRA, H.N.; MAGNABOSCO, C.U.; ZAMBIANCHI, A.R.; ALBUQUERQUE L.G.;

BERGMANN, J.A.G.; SAINZ, R.D. **Avaliação Genética de Touros e Matrizes da Raça Nelore: sumário 2005**. Ribeirão Preto: ANCP, 2004.

LÔBO, R.B. Memorial. Ribeirão Preto: FMRP/USP, 2005.

LÔBO, R.B. **Programa de Melhoramento Genético da Raça Nelore**. Ribeirão Preto: Raysildo Barbosa Lôbo, 1992.

LÔBO, R.B.; BEZERRA, L.A.F.; OLIVEIRA, H.N.; MAGNABOSCO, C.U.; ZAMBIANCHI, A.R.; ALBUQUERQUE L.G.; BERGMANN, J.A.G.; SAINZ, R.D. **Avaliação Genética de Touros e Matrizes da Raça Nelore: sumário 2005**. Ribeirão Preto: ANCP, 2005.

LUSH, J.L. **Melhoramento Genético dos Animais Domésticos**. Rio de Janeiro: Centro de Publicações Técnicas da Aliança Norte-Americana de Cooperação Econômica e Técnica no Brasil – USAID, 1964.

MANNILA, H. Data Mining: machine learning, statistic and databases. (Technical Report FIN-0014). Helsinki: University of Helsinki, 1997

MARQUES, V.F. Analisando os Dados do Programa de Melhoramento Genético da Raça Nelore com Data Warehousing e Data Mining 116 p.; 2002. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2002.

MENDONÇA NETO, M.G.; NOGUEIRA, L.A.H.N.; PONTES, L.A.M.; TEIXEIRA, L.S.G.T.; GUIMARÃES, P.R.B. Aplicação de Técnicas de Mineração Visual de Dados na Regulação da Indústria de Energia: um estudo de casos. (Relatórios Técnicos: RT-NUPERC-2000-1/p). Salvador: Universidade Salvador, 2000.

MENDONÇA NETO, M.G.; SUNDERHAFT, N.L. A State-of-the-Art-Report – Mining Software Engineering Data: a survey. Rome: Data & Analysis Center for Software (DACS), Department of Defense (DoD), 1999.

MILLER, G.A. The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. **Psychological Review**, Washington, v. 101, n. 2, p. 343-352, 1994.

MONARD, M.C.; BATISTA, G.E.A.P.A; KAWAMOTO, S.; PUGLIESE, J.B. Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos. (Technical Reports). São Carlos: ICMC-USP, 1997.

OLIVEIRA, H.N. Grupos de Contemporâneos e Conectabilidade. In: WORKSHOP SELEÇÃO EM BOVINOS DE CORTE, S., 2003, Salvador. **Anais...** Ribeirão Preto: ANCP, 2003. CD-ROM.

ORACLE. Oracle 8i: data warehouse guide. (Release 2 (8.1.6)). Redwood Shores: Oracle Corporation, 1999.

PEREIRA, J.C.C. **Melhoramento Genético Aplicado à Produção Animal**. 4ª ed. Belo Horizonte: FEPMVZ, 2004.

PIATETSKY-SHAPIRO, G. Knowledge Discovery in Real Databases. **AI Magazine**, Menlo Park, v. 11, n. 5, p. 68-70, 1991.

RAZENTE, H.L. Análises Visuais em Processos de Redução de Dimensionalidade para Mineração em Sistemas de Base de Dados 69 p.; 2004. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

REZENDE, S.O. Introdução. In: REZENDE, S.O. **Sistemas Inteligentes: fundamentos e aplicações**. Barueri: Manole, 2003. p. 3-11.

REZENDE, S.O.; PUGLIESI, J.B.; MELANDA, E.A.; PAULA, M.F. Mineração de Dados. In: REZENDE, S.O. **Sistemas Inteligentes: fundamentos e aplicações**. Barueri: Manole, 2003. p. 307-335.

ROHRER, R.M.; SIBERT, J.L.; EBERT, D.S. A Shape-Based Visual Interface for Text Retrieval. **IEEE Computer Graphics and Applications**, Los Alamitos, v. 19, n. 5, p. 40-46, 1999.

SAS. SAS 9.1.3. Cary: SAS Institute Inc., 2003.

SERVIÇO DE INFORMAÇÃO DA CARNE [SIC]. Nutrição. São Paulo. Disponível em: http://www.sic.org.br/. Acesso em: 09 fev. 2006b.

SERVIÇO DE INFORMAÇÃO DA CARNE [SIC]. Você Sabe para que Serve um Boi? São Paulo. Disponível em: http://www.sic.org.br/. Acesso em: 09 fev. 2006a.

SHIMABUKURO, M.H. Visualizações Temporais em uma Plataforma de Software Extensível e Adaptável 147 p.; 2004. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

SHNEIDERMAN, B. Dynamic Queries for Visual Information Seeking. **IEEE Software Magazine**, Los Alamitos, v. 6, n. 11, p. 70-77, 1994.

SHOSHANI, A. OLAP and Statistical Database: similarities and differences. **ACM TODS**, New York, p. 185-87, 1997.

SILBERSCHATZ, A.; TUZHILIN, A. On Subjective Measures of Interestingness in Knowledge Discovery. In: ICML WORKSHOP ON APPLYING MACHINE LEARNING IN PRACTICE, S., 1995, **Anais…** Tahoe City: D. Aha & P. Riddle, 1995. p. 50-56.

SPOTFIRE. Spotfire DecisionSite 6.0.0. Cambridge: Spotfire Inc., 2000.

TONHATI, H.; MARCONDES, C.R.; LÔBO, R.B. Sumários e Aplicações. In: WORKSHOP SELEÇÃO EM BOVINOS DE CORTE, S., 2003, Salvador. **Anais...** Ribeirão Preto: ANCP, 2003. CD-ROM.

VAN VLECK, L.D. Contemporany Groups for Genetic Evaluations. **J. Dairy Science**, Savoy, v. 70, p. 2456-2464, 1987.

VAN VLECK, L.D. **Selection Index and Introduction to Mixed Model Methods for Genetic Improvement of Animals: the green book**. Florida: CRC Press, 1993.

VOZZI, P.A. Análise da Estrutura e Variabilidade Genética dos Rebanhos do Programa de Melhoramento Genético da Raça Nelore 77 p.; 2004. Dissertação (Mestrado) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2004.

WIKIPEDIA. OLE DB. Disponível em: http://en.wikipedia.org/. Acesso em: 20 abr. 2006.

WOLPERT, D.H. Stacked Generalization. **Neural Networks**, London, v. 5, n. 2, p. 241-259, 1992.

WONG, P.C. Visual Data Mining. **IEEE Computer Graphics and Applications**, Los Alamitos, v. 19, n. 5, p. 20-21, 1999.

WONG, P.C.; BERGERON, R.D. 30 Years of Multidimensional Multivariate Visualization. Scientific Visualization – Overviews, Methodologies and Techniques, **IEEE Computer Society Press**, Los Alamitos, p. 3-33, 1997.