ÍTALO TSUCHIYA

REGRESSÃO LOGÍSTICA APLICADA NA ANÁLISE ESPACIAL DE DADOS ARQUEOLÓGICOS

Dissertação apresentada ao Curso de Pós – Graduação em Ciências Cartográficas para a obtenção do Título de Mestre em Ciências Cartográficas pela Universidade Estadual Paulista.

PRESIDENTE PRUDENTE

2002

ÍTALO TSUCHIYA

REGRESSÃO LOGÍSTICA APLICADA NA ANÁLISE ESPACIAL DE DADOS ARQUEOLÓGICOS

Dissertação apresentada ao Curso de Pós – Graduação em Ciências Cartográficas para a obtenção do Título de Mestre em Ciências Cartográficas pela Universidade Estadual Paulista.

Orientadora
Profa. Dra. Vilma Mayumi Tachibana

Co-orientador
Prof. Dr. Nilton Nobuhiro Imai

PRESIDENTE PRUDENTE

2002

REGRESSÃO LOGÍSTICA APLICADA NA ANÁLISE ESPACIAL DE DADOS ARQUEOLÓGICOS

por

Ítalo Tsuchiya
Engenheiro Cartógrafo

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Ciências Cartográficas da Universidade Estadual Paulista, pela comissão formada pelos professores.

Prof. Dr. José Alberto Quintanilha

Profa. Dra. Rosângela Custódio Cortez Thomaz

Profa. Dra. Vilma Mayumi Tachibana

Presidente Prudente, dezembro de 2002.

DEDICATÓRIA

Dedico esse trabalho aos meus Pais,
e aos meus irmãos

AGRADECIMENTOS

“Primeiramente ao Deus, pois quando Ele quer não tem quem não queira”

Aos que me orientaram e auxiliaram muito nesse trabalho: Vilma Mayumi Tachibana e Nilton Nobuhiro Imai

À banca examinadora: José Alberto Quintanilha e à Rosângela Custódio Cortez Thomaz.

Aos professores da graduação em Engenharia Cartográfica e Pós em Ciências Cartográficas principalmente a Mauro I. Ishikawa, Otávio Y. Itame, Erivaldo A. da Silva, José M. Arana, Maria de Lourdes Galo, Tereza H. Yamabe, Maurício Galo, Paulo de Oliveira e Messias Meneguette Jr.

Aos Departamentos de Cartografia e Planejamento, principalmente a Graça, Cidinha, Soninha, Ruth e Leonice.

A todos os funcionários da FCT/Unesp, especialmente Geraldo, Gege, Milton, Flora, Maria, Walmir, Sílvio Tadao Fujisaki, Cláudia, Fátima, Marisa, Raquel, Geny, Antério, Pedro, Moacir, Eduardo, Edmilson, Mauro, Sávio, Geraldinho, Nilson, Donato, Caldeira, Bubu, Francisco, Paulo Ruiz, Fátima Paulino, Cida e as pessoas a qual não me lembrei, mas que estão sempre presentes.

Aos amigos Marcelo A. Cirillo, Luis F. Sapucci, Daniel R. dos Santos, Kátia L. Oliveira, Lucinha, Fábio L. de Almeida, Wagner Carrupt, Eduardo A. Silva, João Bosco N. Jr., André Castro, José L. Maia, Elivagner B. de Oliveira, France M. Ferreira, João Osvaldo, Elias A. Jr, Giovane M. do Vale e José A. Perez.

Aos amigos da empresa Sementes Selecta, especialmente à Telma Mendes, Hélio Okumoto, Ademir Honório, Cátia Vaz de Avila, e Tininha.

Aos amigos do Programa de Pós Graduação em Ciências Cartográficas e Graduação Engenharia Cartográfica.

Aos amigos da Sala 12, Sala 20, Latogeo, Laboratório de Foto, Laboratório Móvel, Estação GPS e Estação Metereológica

“Para muitas questões matemáticas há uma resposta definida.
Esta pode ser muito difícil encontrar, abrangendo muitos problemas técnicos,
e poderemos ter de nos contentar com uma aproximação”

Meyer, 1978, referindo-se à Estimação de Parâmetros.

SUMÁRIO

TÍTULO i
TERMO DE APROVAÇÃO iii
DEDICATÓRIA iv
AGRADECIMENTOS v
EPÍGRAFE vi
SUMÁRIO vii
LISTA DE FIGURAS ix
LISTA DE TABELAS x
RESUMO xi
ABSTRACT xi

1. INTRODUÇÃO 1
1.1. Objetivos 2
1.2. Conteúdo do Trabalho 3

2. GEOPROCESSAMENTO 4
2.1. Sistema de Informação Geográfica 4
2.1.1. Modelagem de dados 6
2.1.2. Modelos tipo geo-campo 8
2.1.3. Representação dos geo-campos 11
2.2. Álgebra de Mapas 13

3. ANÁLISE DE DADOS ESPACIAIS EM ARQUEOLOGIA 16
3.1. Modelagem preditiva 17
3.1.1. Método por decisão por árvore 20
3.1.2. Modelagem através de decisão por árvore 22
3.2. Análise discriminante pelo método de Fisher 25
3.2.1. Função discriminante de Fischer para 2 populações 26
3.2.2. Problema de classificação geral 30

4. REGRESSÃO LOGÍSTICA 35
4.1. Modelo de regressão logística 36
4.1.1. Estimativas de máxima verossimilhança 41
4.1.2. Teste de significância dos coeficientes 44

5. MATERIAL E MÉTODO 50
5.1. Material 50
5.2. Modelagem dos dados arqueológicos 51
5.2.1. Problemática 52
5.2.2. Modelo Estático 53
5.2.3. Universo de representação 55
5.3. Banco de dados geográficos (BDGs) 56
5.4. Ajustando o modelo de regressão logística 57
5.5. Utilizando o modelo de regressão logística 58

6. RESULTADO 60
6.1. Banco de Dados Geográficos 60
6.2. Ajustando o modelo de regressão logística 73

7. CONCLUSÃO 82
7.1. Recomendações 85

BIBLIOGRAFIA

LISTA DAS FIGURAS

Figura 01 05
Figura 02 06
Figura 03 06
Figura 04 07
Figura 05 09
Figura 06 12
Figura 07 18
Figura 08 19
Figura 09 21
Figura 10 24
Figura 11 28
Figura 12 31
Figura 13 37
Figura 14 38
Figura 15 50
Figura 16 55
Figura 17 56
Figura 18 58
Figura 19 59
Figura 20 64
Figura 21 65
Figura 22 66
Figura 23 67
Figura 24 68
Figura 25 69
Figura 26 70
Figura 27 71
Figura 28 72
Figura 29 78
Figura 30 80
Figura 31 82
Figura 32 84

LISTA DAS TABELAS

Tabela 01 32
Tabela 02 37
Tabela 03 62
Tabela 04 63
Tabela 05 74
Tabela 06 75
Tabela 07 76
Tabela 08 76

RESUMO - O presente trabalho tem como referência o Projeto de Salvamento Arqueológico de Porto Primavera, cujo objetivo foi resgatar, analisar e conservar os vestígios das antigas civilizações ribeirinhas do Rio Paraná. Uma das etapas de um projeto de salvamento é a prospecção, nela, o arqueólogo realiza a vistoria em toda a área de interesse à procura de indícios das habitações pretéritas, porém, essa etapa é demorada e onerosa. Com base na técnica de modelagem preditiva multivariada (regressão logística), aliada às ferramentas de Geoprocessamento, pudemos avaliar as prováveis áreas de ocupação pretérita, reduzindo assim, a etapa de prospeção. Como resultado temos um mapa de classes, com as probabilidades de ocorrência de sítios arqueológicos utilizando o método de regressão logística.
Palavras chave: **Geoprocessamento, Análise Espacial, Regressão Logística, Modelos Preditivos e Arqueologia.**

ABSTRACT – This work has as reference the Project of Archaeological Rescue of “Porto Primavera”, whose objective was to rescue, to analyze and to conserve the vestiges of the old marginal civilizations of the Paraná River. One of the stages of project is the prospection, in, the archaeologist carries through the inspection in all the area of interest to the search of indications of the past habitations, however, this stage is delayed and onerous. On the basis of the technique of multivaried predictive modeling (logistic regression), allied to the tools of Geographic Information System, we could evaluate the probable areas of past occupation, thus reducing, the stage of prospection. As result we have a map of classrooms, with the probabilities of occurrence of archaeological small farms using the method of logistic regression.
Keywords: **Geomatic, Spatial Analysis, Logistic Regression, Preditction Models and Archaeology.**

1. INTRODUÇÃO

Este trabalho tem como referência o Projeto de Salvamento Arqueológico de Porto Primavera - SP/MS, cujo objetivo foi resgatar o patrimônio arqueológico existente na área de influência da formação do lago da Usina Hidrelétrica Engenheiro Sérgio Motta (Porto Primavera).

Esse tipo de resgate é denominado Arqueologia de Salvamento ou de Contrato, sendo necessário em grandes empreendimentos como gasodutos, projetos de urbanização, hidrelétricas, entre outros.

A Arqueologia de Contrato é realizada por meio de convênios, onde a empresa contratante oferece recursos financeiros para a equipe ou instituição, que realiza o resgate do material arqueológico, que possa estar sob a área abrangida pelo empreendimento.

Segundo Caldarelli (1989), a exemplo do que ocorreu em outros países, a pesquisa arqueológica levada a cabo no Brasil, é predominantemente realizada por contrato de prestação de serviços, no qual os arqueólogos elaboram pareceres técnicos, integrados ao Estudo de Impacto Ambiental (EIA) e o Relatório de Impacto do Meio Ambiente (RIMA), diferenciando-se da Arqueologia Acadêmica, cujo objetivo é o crescimento teórico da Ciência.

Um Projeto de Salvamento Arqueológico divide-se nas etapas de campo, gabinete e laboratório, sendo que na primeira são realizadas as prospecções para a procura de indícios de habitações ou passagens dos índios e a escavação para o resgate desse material. Na segunda parte são elaborados os relatórios, que serão entregues à empresa contratante. Na última etapa são realizadas análises dos materiais coletados, bem como o arquivamento das informações no acervo arqueológico.

O foco deste projeto está na prospecção, que é a etapa na qual os arqueólogos verificam a existência de ocupações ou passagens de antigas civilizações em campo. Esse processo é realizado pelos pesquisadores com base no conhecimento de como se comportavam as populações, bem como a associação das variáveis ambientais como geomorfologia, geologia, altimetria e vegetação.

Na prospecção faz-se o caminhamento em toda área à procura de vestígios arqueológicos.

Para Morais (1990), a prospecção é um processo inicial, determinado como reconhecimento da área, durante o qual são realizadas missões de averiguação do terreno, com base em informações cartográficas, textos especializados e declarações da comunidade local. Nesse processo são demarcadas as áreas onde estão localizados os sítios e ocorrências arqueológicas, sendo que o primeiro denota grande presença de material e o segundo menos material, deduzindo-se como área de moradia e acampamento, respectivamente. A verificação de sua existência no campo permite o posicionamento através da determinação das coordenadas geográficas e o registro no material cartográfico existente. No decorrer desta verificação são, também, feitas coletas de material arqueológico comprobatório, com o objetivo de garantir o testemunho e possibilitar uma análise mais acurada, em laboratório, da potencialidade em vistas de uma futura escavação do sítio.

Como visto acima, o processo de prospecção é demorado e oneroso, pois trata de uma busca de material das antigas ocupações em campo. Assim sendo, propõe-se verificar a performance da análise espacial (regressão logística) em um modelo de dados geográficos do lago de inundação da UHE Sérgio Motta, a fim de gerar um produto adequado à predição de locais mais propícios para conter indícios de antigas civilizações.

Espera-se, com isso, contribuir para a otimização do processo de prospecção em Projetos de Salvamento Arqueológico e pesquisas arqueológicas.

1.2. Objetivos

O presente trabalho objetiva determinar parâmetros de um modelo matemático que represente a probabilidade de ocorrência de um evento de interesse, tal como presença de sítios arqueológicos, baseado nas variáveis ambientais como geologia, geomorfologia, pedologia e altimetria e, em alguns locais conhecidos, de presença ou ausência de vestígios arqueológicos (dicotômicos).

A determinação dos parâmetros dar-se-á pelo método de Regressão Logística, que seleciona as variáveis mais relevantes ao modelo arqueológico da população ribeirinha que habitava o Rio Paraná, podendo, a partir do modelo ajustado, predizer as prováveis áreas de ocupações pretéritas.

Todo o processo poderá validar uma ferramenta de análise espacial para a geração de mapa com as prováveis áreas de ocupações pretéritas, otimizando uma das etapas do Projeto de Salvamento Arqueológico: a prospecção.

1.3. Conteúdo do trabalho

Apresenta-se o conteúdo da pesquisa em sete capítulos: o primeiro constitui-se de uma breve introdução e os objetivos do trabalho, onde é abordada sua importância para a Arqueologia e seus processos.

No segundo capítulo são apresentados conceitos de Sistema de Informação Geográfica (SIG), com enfoque sobre modelos do tipo geo-campo, utilizando-se o conceito dos universos de abstração adaptado por Câmara et. al.(1996) para o caso de SIG.

No terceiro capítulo são apresentados alguns tipos de análises espaciais para predição de sítios arqueológicos, com ênfase no método de classificação por árvore e análise da função discriminante de Fisher.

No quarto capítulo são apresentados os conceitos do modelo logístico, bem como a técnica de teste de hipóteses.

O material e o método da aplicação de Regressão Logística com auxílio das ferramentas de Geoprocessamento são apresentados no capítulo cinco.

Os resultados obtidos a partir da Regressão Logística para o conjunto de dados do Projeto de Porto Primavera são mostrados no capítulo seis.

As conclusões e recomendações do autor compõem o sétimo capítulo desta dissertação.

2. GEOPROCESSAMENTO

A análise espacial, como a que se propõe avaliar neste trabalho, insere-se no contexto do Geoprocessamento, que pode ser considerado como um conjunto de ferramentas de processamento e descrições de informações geográficas. Entre essas ferramentas podem ser citados os métodos geoestatísticos, as ferramentas de Sistema de Informação Geográfica, os processamentos de imagens e os métodos de interpolação.

Segundo Teixeira (1992), Geoprocessamento é uma tecnologia que abrange o conjunto de procedimentos de entrada, manipulação, armazenamento e análise de dados espacialmente *georreferenciados*.

Para Câmara (1996), Geoprocessamento denota a disciplina do conhecimento que utiliza técnicas matemáticas e computacionais para o tratamento da informação geográfica e que vem influenciando de maneira crescente as áreas de Cartografia, Análise de Recursos Naturais, Transportes, Comunicações, Energia e Planejamento Urbano e Regional. As ferramentas computacionais para o Geoprocessamento, chamadas de Sistemas de Informação Geográfica, permitem realizar análises complexas, ao integrar dados de diversas fontes e ao criar bancos de dados *georreferenciados*. Tornam ainda possível automatizar a produção de documentos cartográficos.

Baseado nesse conhecimento apresentam-se os conceitos básicos de um Sistema de Informação Geográfica.

2.1. Sistema de Informação Geográfica

O Sistema de Informação Geográfica, a princípio, é uma especialização dos Sistemas de Informação (SI) que, segundo Laudon (1999), é definido como um conjunto de componentes inter-relacionados trabalhando junto para coletar, recuperar, processar, armazenar e distribuir informação com a finalidade de facilitar o planejamento, o controle, a coordenação, a análise e o processo decisório em empresas e outras organizações.

Um Sistema de Informação contém informações sobre pessoas, lugares e coisas de interesse no ambiente ao redor da organização e dentro da própria organização, transformando a informação em uma forma utilizável para a coordenação de fluxo de trabalho

de uma empresa, ajudando empregados ou gerentes na tomada de decisões, análises, visualizações de assuntos complexos e na resolução de outros tipos de problemas.

Num SI existe um ciclo de três atividades básicas: entrada, processamento e saída, como pode ser observado na Figura 01.

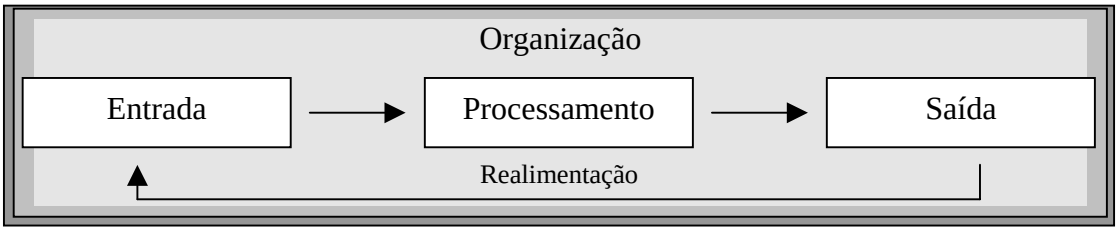


Figura 01: Transformação dos dados originais em informação útil (Fonte: Laudon, 1999)

Um Sistema de Informação Geográfica é um tipo de SI e surgiu da necessidade de armazenamento, processamento e visualização de dados espaciais, juntamente com o alto desenvolvimento computacional.

Worboys (1995) define o SIG como um Sistema de Informação baseado em computador, que permite capturar, modelar, manipular, corrigir, analisar e apresentar dados *georreferenciados*. A tecnologia permite o monitoramento de eventos, fortalecendo o processo de planejamento e organização de qualquer sistema geográfico informal. Basicamente, a informação é tratada sob a forma de base de dados *georreferenciados*, mantendo uma associação com as informações gráficas tais como pontos, linhas e polígonos e, com informações não espaciais (exemplo: a iluminação do poste é “amarela”).

Para Campbel (1995), SIG é, na sua essência, um conjunto de tecnologias baseado em computador que são hábeis para armazenar, visualizar, manipular e analisar dados espaciais, mais particularmente mapas baseados em informação.

E para Câmara (1996), o termo Sistema de Informação Geográfica é aplicado para sistemas que realizam o tratamento computacional de dados geográficos. Um SIG armazena a geometria e os atributos dos dados que estão *georreferenciados*, isto é, localizados na superfície terrestre e representados numa projeção cartográfica. Os dados tratados em Geoprocessamento têm como principal característica diversidade de fontes geradoras e de formatos apresentados.

Todas as definições descrevem o Sistema de Informação Geográfica como ferramenta de análise, visualização e armazenamento de dados geográficos, utilizando

ferramentas computacionais e recursos humanos. Essas ferramentas são utilizadas no auxílio de um determinado processo, provenientes da abstração ou simplificação do mundo real, como mostra a Figura 02.

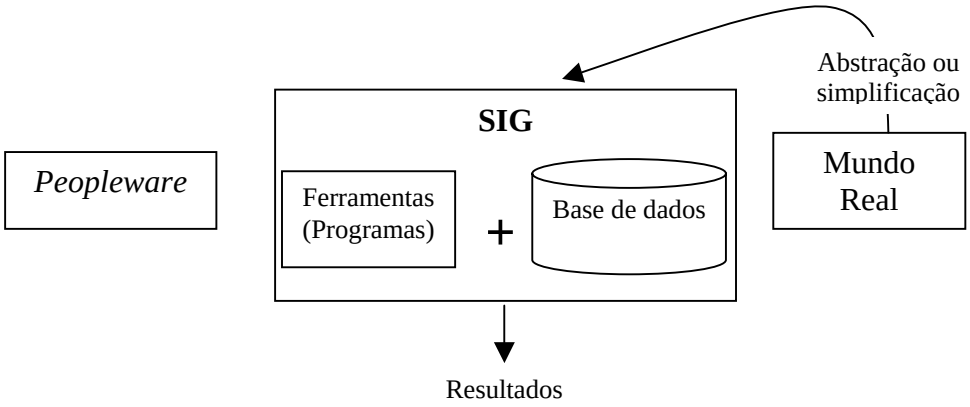


Figura 02: Componentes de um Sistema de Informação Geográfica (Fonte: ESRI, 1995)

Mais detalhes sobre Sistema de Informação Geográfica podem ser encontrados em Câmara et. al. (2002), Worboys (1995), Campbell (1995), entre outros.

2.1.1. Modelagem de dados

Na modelagem de um Sistema de Informação Geográfica deve-se levar em consideração as componentes que o envolvem, não somente analisar qual programa utilizar, mas também ter a preocupação da integração das três componentes básicas: organizações, pessoas e a tecnologia, para que resulte em um trabalho coeso e eficiente.

Um SIG utiliza o esquema dos quatro universos no processo de abstração, como apresentado na Figura 03.

Câmara (1996) descreve o paradigma desses quatro universos voltados para Sistemas de Informação Geográfica.

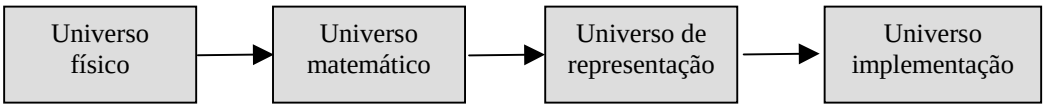


Figura 03: Níveis conceituais de abstração (Fonte: Câmara, 1996)

Segundo Borges (1997), o modelo busca sistematizar o entendimento que é desenvolvido a respeito de objetos e fenômenos que serão representados em um sistema informatizado. Os objetos e fenômenos reais, no entanto, são complexos demais para permitir uma representação completa, considerando os recursos à disposição dos sistemas gerenciadores de bancos de dados (SGBD) atuais. Desta forma, é necessário realizar uma abstração dos objetos e fenômenos do mundo real, de modo a se obter uma forma de representação conveniente, embora simplificada, que seja adequada às finalidades das aplicações do banco de dados.

A modelagem de um Sistema de Informação Geográfica requer um conhecimento e uma habilidade prática com tecnologias de informação geográfica; uma compreensão de organizações e indivíduos, com uma perspectiva comportamental (relacionamento e funcionalidade) e uma compreensão ampla de como analisar e resolver problemas para os usuários, como mostra a Figura 04.

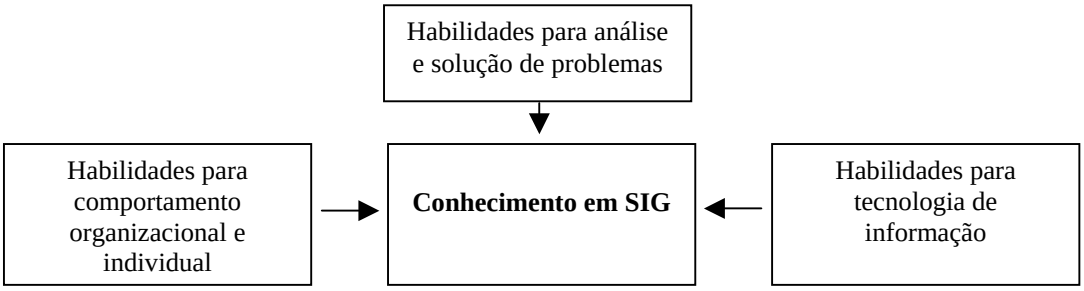


Figura 04: Temas centrais do conhecimento em Sistemas de Informação (Laudon, 1999).

Assim a modelagem em SIG trata-se de uma técnica de projeto, que visa a análise dos processos inerentes à realização de determinadas atividades, servindo para diversos objetivos. Rumbaugh (1991) modela um projeto na seguinte seqüência:

- Testar uma entidade física antes de lhe dar a forma;
- comunicação com clientes (arqueólogos);
- visualização e
- redução da complexidade.

Em resumo, utiliza-se a modelagem, pois os computadores e, conseqüentemente, os Sistemas de Informação Geográfica operam sobre números e caracteres e não há como aplicá-los diretamente às variáveis do mundo real. A representação e a análise

destas variáveis ambientais no contexto computacional passam primeiramente pela aquisição de amostras da variável. Contudo, devido à complexidade da natureza, à imprecisão na medida das amostras, às aproximações de modelagem, aos interesses de estudo e às limitações computacionais, os modelos são representações simplificadas da realidade. Um bom modelo é aquele que prevê correta e consistentemente o funcionamento do mundo real para a variável de interesse (Barbosa, 1997).

A modelagem do mundo real num banco de dados geográficos é uma atividade complexa, pois como já foi colocado, envolve a discretização do espaço geográfico (Davis, 2002). Nessa obra são apresentados os seguintes fatores envolvidos no processo de discretização do espaço: transição da informação em unidades lógicas de dados, forma como as pessoas percebem o espaço, natureza diversificada dos dados geográficos, existência das relações espaciais, coexistência de entidades essenciais ao processamento e atividades "cartográficas".

No presente trabalho, como os dados referem-se, principalmente, a fenômenos que variam continuamente no espaço (como geologia, geomorfologia, altimetria, ...), a visão de campos (Worboys, 1995) é adotada para a representação adequada desses fenômenos.

Assim, as representações do Mundo Arqueológico (Domínio Fonte) são funções do tipo f(E,N), onde E e N são as coordenadas referentes à área geográfica do projeto (domínio espacial) e essas funções têm como contradomínio os valores associados às variáveis ambientais (independentes).

Mais informações sobre geo-objetos podem ser obtidos em Worboys (1995), Câmara et. al. (2002) e Barbosa (1997).

2.1.2. Modelos tipo geo-campo

Segundo Worboys (1995), o modelo de geocampo trata a informação como um conjunto de distribuições espaciais, onde cada distribuição pode ser formalizada como uma função matemática de uma área geográfica, representando uma distribuição geográfica contínua. A Figura 05 representa as localizações de uma área geográfica baseada em geo-campos para *n* variáveis.

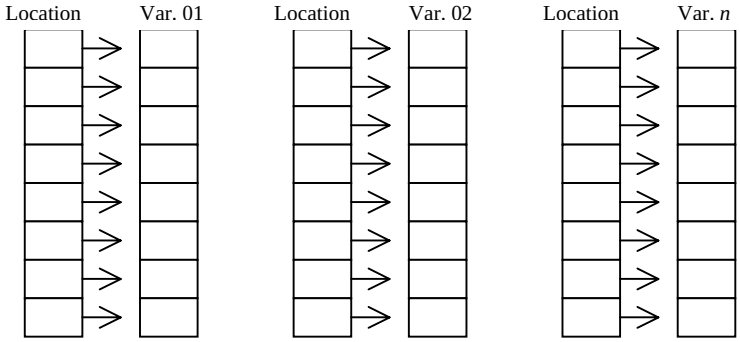


Figura 05: Fenômenos geográficos baseados em geo-campo (Fonte: Worboys, 1995)

Para Câmara (1996), o geo-campo representa a distribuição espacial de uma variável que possui valores em todos os pontos pertencentes a uma região geográfica, na qual um *geo-campo* (*f*) é uma entidade matemática que representa a distribuição de uma variável espacialmente contínua sobre uma região geográfica (*R*):

$$f = [R,V,\lambda],$$

onde:

R : região geográfica definindo o domínio espacial
V : contra-domínio de valores da variável na região geográfica
λ: mapeamento entre pontos (x,y) em R e valores em V ($\lambda : R \rightarrow V$)

Um modelo de geo-campo consiste em uma coleção finita de *n* campos espaciais, $\{f_i : 1 \leq i \leq n\}$. Para $1 \leq i \leq n$, cada campo espacial f_i é uma função computável de um conjunto F para um atributo finito do domínio A_i.

Para uma melhor compreensão dos geo-campos é necessário entender as suas propriedades, que podem ser de estrutura espacial e domínio de atributos (Worboys, 1995).

O domínio de atributos pode conter valores mensuráveis pertencentes aos seguintes tipos (Worboys, 1995):

a) nominais: valores qualitativos em que não podem ser aplicadas as operações aritméticas. Por exemplo, o nome de um tipo de vegetação;
b) ordinais: quantidades por ordenação numa escala linear, mas não por magnitude. Valores ordinais podem ser comparados por tamanho, mas não podem ser adicionados, multiplicados, subtraídos, etc. Como exemplo, o nível de escolaridade nas regiões;
c) intervalares: quantificados pela posição relativa num intervalo de escala, onde as medidas de intervalo podem ser comparadas por tamanho, com a magnitude da diferença, obtendo-se uma noção expressiva e
d) proporção: ao contrário dos intervalos, é definido com respeito a um ponto fixo, sendo permitidas as operações aritméticas (adição, subtração multiplicação e divisão). Pode-se citar a altitude sobre o nível médio dos mares.

Os geo-campos podem ser especializados em (Câmara, 1996):

- Temático: usado para modelar variáveis cujos valores foram obtidos através de escalas de medidas temáticas, binária, nominal e ordinal. A função f caracteriza um geo-campo temático quando define um mapeamento $f : R \rightarrow V$, tal que V é um conjunto finito enumerável. Os elementos de V definem os temas de um geo-campo temático (p.ex. um mapa de solos é caracterizado pelo conjunto de temas {latosolo roxo, litosolo, podzólico vermelho, etc.});

- Numérico: usado para modelar variáveis cujos valores foram obtidos através de escalas de medidas numéricas, intervalares e proporcionais. A função f caracteriza um geo-campo numérico quando define um mapeamento $f : R \rightarrow V$, tal que V é o conjunto dos reais. Dados de altimetria, temperatura, pressão são exemplos de dados geográficos que podem ser modelados conceitualmente como geo-campo;

- Imagem: usado para modelar variáveis cujos valores foram obtidos através de discretização da resposta recebida por sensor remoto (passivo ou ativo) de uma área da superfície terrestre. A função f caracteriza um geo-campo imagem quando define um mapeamento $f : R \rightarrow V$, tal que V é o conjunto dos naturais. Esta classe é uma especialização da classe geo-campo numérico.

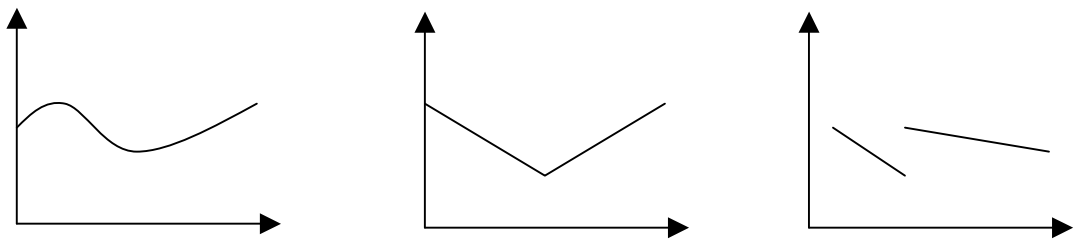
Cliff e Ord (1981) apud Barbosa (1997) diz que uma característica particular dos campos físicos é o grau extremamente alto de autocorrelação espacial. Essa característica permite inferir que posições próximas têm chance de possuírem características similares, compensando assim o fato de não se medir o fenômeno contínuo em todas as posições, dado que elas são infinitas.

2.1.3. Representação dos geo-campos

Após a modelagem conceitual deve ser gerado o modelo de representação, onde a geometria é representada no computador, baseado nos modelos matemáticos definidos no modelo conceitual.

Como os geo-campos representam dados espaciais do Mundo Real como função do tipo f(x,y), esses podem ser representados, segundo Worboys (1995), como os seguintes tipos:

- Contínuo: no qual havendo uma pequena variação na localização geográfica haverá também uma pequena mudança no valor do atributo, não ocorrendo variações abruptas num intervalo pequeno, ou seja, a forma desse tipo de geo-campo é uma função suavizada (Figura 06a);
- Diferenciáveis: nesse tipo de geo-campo, a taxa de mudança é bem definida (declive), como mostra a Figura 06b;
- Discreto: nesse geo-campo pode-se notar que graficamente, há uma descontinuidade na curva, ou seja, os limites para as classes são bem definidos (Figura 06c).



(a) (b) (c)

Figura 06: Geo-campos do tipo (a) contínuo, (b) diferenciado e (c) discreto (Worboys, 1995)

Esses tipos de representação são modelados de acordo com problema a ser aplicado pelo usuário, pois cada abstração do Mundo Real é particular à determinada aplicação. Neste trabalho utilizou-se o programa Arc/Info, que tem os permite gerar os seguintes tipos de geo-campo (Barbosa, 1997):

- *Grid* ou *Lattice*: são grades regulares com determinado limite geográfico, resolução e valores da variável associados, como uma matriz de células, onde para cada posição (x,y) tem-se um valor associado. Esses valores podem ser do tipo *float* (real) ou *integer* (inteiro);

- *Regions* (polígonos): são representações poligonais regulares ou irregulares, que têm valores atribuídos a cada polígono. Cada região tem um valor E, N (centróide) e um valor numérico, escalar ou dado alfanumérico, como é o caso da cobertura vegetal;

- Isolinhas: são linhas que têm o mesmo valor de variável. Numa determinada área geográfica têm-se várias isolinhas ordenadas, onde, por regra, nunca se cruzam. As isolinhas são formadas por pequenas linhas, para os quais são atribuídas coordenadas de início, passagem, fim e o valor da variável. Como exemplo, as representações de curvas de nível e as linhas isotermas;

- *TIN (Triangular Irregular Network)*: a rede de triângulos irregulares representa um terreno por meio de triângulos irregulares, onde os vértices são pontos amostrados no terreno. Os *TINs* são gerados por

amostras pontuais ou isolinhas, cujos valores associados representam a variável em questão;

- Amostras de pontos irregularmente espaçados: nesse caso, as coordenadas não respeitam um padrão de locação entre as amostras, perfazendo um conjunto de coordenadas E, N irregulares com o valor da variável associado. Como exemplo pode ser citada a coleta de amostras aleatórias de solo em "Z", que atualmente é utilizada pelos técnicos agrícolas;

- Amostra de pontos regularmente espaçados: esse tipo de representação aloca coordenadas E, N e o valor da variável na área geográfica, porém respeitando-se uma eqüidistância entre as amostras. Segundo Barbosa (1997), o espaçamento entre as amostras deve ser no mínimo duas vezes menor que a menor variação espacial presente no fenômeno em estudo, ou a freqüência espacial da amostra deve ser duas vezes maior que a maior freqüência espacial presente no fenômeno em estudo.

2.2. Álgebra de Mapas

A álgebra de mapas é definida como uma linguagem para manipulação das representações cujo domínio é comum, ou seja, os planos de informação compreendem a mesma região geográfica.

Como o presente trabalho trata da determinação de um modelo logístico para a predição de prováveis áreas de ocupação pretérita e sua aplicação em uma região geográfica, o projeto envolve a realização de uma seqüência de operações algébricas com os diversos planos de informação após o ajuste do modelo preditivo. Essas operações são realizadas com recursos acessados através de uma linguagem de manipulação.

Nesse tópico é dada ênfase à álgebra de campos utilizando o módulo Arc*Grid* do Arc/Info, que provê ferramentas de análises para *grids* (representações matriciais).

Segundo Berry apud Barbosa (1997), o conceito de álgebra de mapas ou álgebra de campos pode ser visto como uma extensão da álgebra tradicional, com um conjunto de operadores onde as variáveis manipuladas são campos geográficos. Estes operadores manipulam um, dois ou mais geo-campos, em que cada geo-campo descreve um atributo diferente ou até um mesmo atributo, mas com valores diferentes.

A base ou elemento atômico desse tipo de processamento é a célula, que pode ser considerada como uma unidade discreta dentro de um domínio **R** (região geográfica).

O processamento de dados espaciais por representações matriciais é uma poderosa ferramenta analítica, pois permite quatro tipos de funções:

- Local (por célula): o valor resultante de cada célula é uma função do valor correspondente à sua localização;
- Focal (por vizinhança): o valor de locação da célula é uma função de valores das células de entrada, especificada pelos seus vizinhos;
- Zonal (por zonas ou valores iguais): o valor resultante da análise leva em consideração células que tenham valores similares ou da mesma categoria, determinadas como zonas;
- Global (por *grid*): o valor de saída de cada locação é potencialmente uma função de todas as células do *grid* de entrada.

Na estrutura interna desse tipo de representação, cada *grid* categórico tem um valor associado numa tabela (VAT – *Value Atribute Table*), que armazena a base de dados relacionais e é análogo ao PAT (*Polygon Attribute Table*) para feições do tipo polígono. Uma VAT é gerenciada por um sistema de base de dados INFO, onde o *grid* é dividido em unidades quadradas denominadas por "*tiles*" e cada um representa uma porção do espaço (Esri, 1995).

Um *tile* forma blocos retangulares, onde há mais blocos no eixo Y do que no eixo X. Essa estrutura foi definida para suportar acesso aleatório aos dados, onde há rápida resposta no caso de uma localização de uma determinada célula. Uma comparação análoga pode ser feita com a carta ao milionésimo do IBGE, onde se tem divisões e subdivisões do Brasil baseados num índice que facilita a localização de uma determinada carta.

Uma outra finalidade para esse tipo de estrutura é a compressão da *coverage* matricial (Plano de Informação matricial), onde o *ArcGrid* decide pelo melhor método de compressão devido ao tipo de dados representados, otimizando o projeto do usuário.

Segundo Cordeiro (1996), a linguagem GRID, embora muito flexível, tem o sério inconveniente de não distinguir entre os diferentes tipos de operandos. Isto decorre do fato de estar ligada às estruturas de dados (matrizes) e não levar em conta a semântica das operações. A linguagem LEGAL do SPRING permite as operações onde são analisados atributos qualitativos e quantitativos, cujos detalhes podem ser obtidos em http://www.dpi.inpe.br/geopro/trabalhos/sbsr96_1.pdf.

3. ANÁLISE DE DADOS ESPACIAIS EM ARQUEOLOGIA

Um SIG não é simplesmente um sistema computacional que faz mapas, embora possa criar mapas em diferentes escalas, em diferentes projeções e com diferentes cores, ele é uma ferramenta analítica. A maior vantagem é que esse sistema permite identificar os relacionamentos espaciais entre as feições dos mapas, não armazenando um mapa no sentido convencional, nem uma imagem particular ou uma visão de uma área geográfica e sim armazenando dados, os quais podem ser desenhados numa visão desejada pela demanda de um propósito particular. (ESRI, 1995).

Nesse capítulo está introduzida uma breve abordagem sobre alguns métodos de análise espacial para dados arqueológicos, para que, a partir da determinação de um modelo, possa ser gerada uma superfície probabilística da existência de fragmentos arqueológicos. Esse resultado é baseado nos planos de informação das variáveis ambientais (independentes ou explicativas) e da variável resposta (presença/ausência de sítios arqueológicos), que no caso é a presença ou a ausência de vestígios arqueológicos.

O aspecto teórico e aplicado da modelagem preditiva em aplicações arqueológicas é um campo relativamente novo na Arqueologia. Esses estudos foram realizados nos anos 50 e 60, mas só obtiveram destaque após os anos 70 (Warren, 1990), juntamente com o desenvolvimento da tecnologia computacional. Nos anos seguintes foi possível realizar modelagem preditiva em Arqueologia aplicada na Cartografia Digital, através das ferramentas de Sistemas de Informação Geográfica (SIG).

A modelagem preditiva se mostra útil na explicação de fenômenos baseados em fatores culturais, como exemplo, predizer a razão pelas quais antigas civilizações preferiam determinadas localizações para a instalação dos acampamentos. Nesses casos, os modelos de predição podem oferecer medidas das fontes culturais em localidades específicas, predizendo um cenário passado para a fixação das antigas civilizações através de um modelo matemático.

Um modelo preditivo pode predizer as prováveis áreas de ocupações pretéritas, baseando-se nas preferências do *habitat* e em variáveis do universo físico, tornando-se útil na etapa de prospecção de Projetos de Salvamento Arqueológico.

O modelo preditivo associa variáveis mais condizentes com o tipo de ocupação através de métodos estatísticos, perfazendo a escolha mais provável dos povos pré-históricos e suas estratégias para realizar as ocupações.

Com base na literatura de Bona (1992), Johnson e Wichern (1992), Tachibana (2000), Galo (2000) e Warren (1990), enfoca-se neste trabalho os seguintes modelos preditivos: árvore de decisão, análise discriminante de Fisher e Regressão Logística.

3.1. Modelagem preditiva

Antes de dar ênfase nos modelos de predição, são aqui apresentadas duas perspectivas básicas dos métodos preditivos: indutivo e dedutivo.

Os modelos indutivos são processos analíticos de fatos particulares ou de casos individuais de uma hipótese que sugerem uma conclusão geral. Segundo Warren (1990), são padrões compostos ou uniformidades que são detectadas em observações empíricas, ou seja, podem ser obtidas novas conclusões a partir do conjunto amostral, permitindo a geração de novos conhecimentos baseados nas observações. Já os modelos dedutivos são gerados da teoria e são análogos ao tipo de predição que são freqüentemente associados com o método científico e nessa visão, não há a criação de novos conhecimentos e sim, a utilização de um conhecimento prévio.

Com base nesse entendimento, podem ser desenvolvidos modelos gráficos e/ou numéricos, que possibilitem manipular variáveis utilizando-se a interseção e/ou os valores de pesos das variáveis.

Um modelo preditivo é compreendido por um conjunto de testes de hipóteses, que considera o método de interação e os valores dos pesos para o modelo. Essa metodologia, enriquecida com o conhecimento dos arqueólogos, permite aprimorar os modelos preditivos aplicados à Arqueologia.

Segundo Bona (1992), há várias suposições sobre o desenvolvimento da modelagem preditiva. A primeira envolve a suposição de que a escolha do local dos acampamentos dos povos pré-históricos foi influenciada por elementos ambientais, naturais e físicos. A segunda é que essas variáveis ambientais sobreviveram e podem ser representadas por dados atuais. Estes dados podem estar em mapas, monografias ou podem ser coletados no

campo. A terceira suposição considera que as correlações entre locais arqueológicos e as características do ambiente físico/natural, observado por arqueólogos, são fatos que indicam as decisões dos povoamentos pretéritos.

Como um modelo preditivo tenta codificar aspectos do comportamento humano, não se pode esperar que ele seja um modelo simplista ou que possa ser desenvolvido de forma imediata, sem muito esforço.

O tempo para o desenvolvimento de um modelo preditivo não é finito. Após a obtenção de um modelo inicial, as observações discrepantes, ou anomalias, devem ser identificadas e serem objeto de uma pesquisa adicional. Se algum padrão for encontrado, muitas anomalias tornam-se previsíveis e elas crescem em importância, necessitando-se de uma nova modelagem. Esta, novamente, tem que ser refinada através das observações que a modelaram, formando um ciclo interminável.

Portanto, a modelagem deve ser considerada e realizada como um processo dinâmico, de modo que qualquer conjunto de dados coletados, em qualquer momento, possa ser incorporado no processo de modelagem para aumentar sua integridade, acurácia e escopo. Considerando também a modelagem preditiva como um processo dinâmico, Bona (1992) descreve-a em três estágios (Figura 07):

- Coleta dos dados e sua organização;
- Desenvolvimento de um modelo inicial e teste deste modelo;
- Introdução de novas variáveis no modelo para refinamentos e aperfeiçoamento.

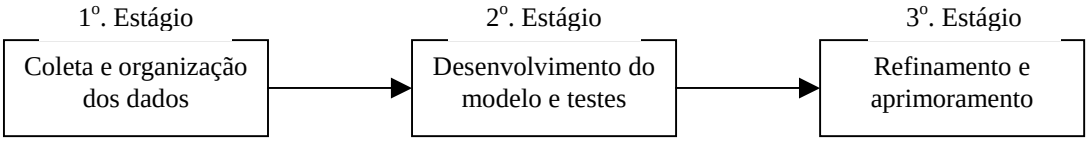


Figura 07: Estágios da modelagem preditiva

No desenvolvimento do primeiro estágio do modelo preditivo, devem ser consideradas:

- (1) As construções das hipóteses que se deseja verificar: Nesse passo, devem ser consideradas as características e atividades dos povos que habitavam a área de pesquisa, pois servem como indicadores

importantes para as variáveis a serem modeladas, ou seja, a maneira que essas variáveis contribuirão para a determinação das prováveis áreas de ocupação pretérita;

- (2) Estratégia para levantamento do conjunto inicial de dados: A partir das hipóteses geradas, tem-se um conjunto de dados iniciais que são relevantes à explicação das hipóteses levantadas no primeiro passo e
- (3) Reconhecimento em campo: Para a validação das hipóteses é necessário que um conjunto de dados apresente observações válidas e seja complementado com as informações sobre eventos já conhecidos. Para tal, é necessário fazer um levantamento de campo.

A figura 08 ilustra os passos do primeiro estágio

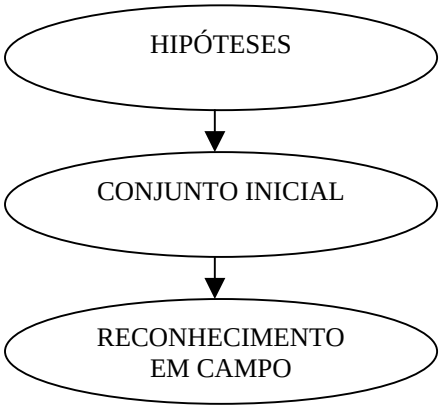


Figura 08: Passos da organização dos dados para modelos preditivos

No desenvolvimento do segundo estágio, também sugerido por Bona (1992), é necessário processar as variáveis organizadas no primeiro estágio. Nesse ponto, as variáveis são introduzidas ou removidas do processo e os pesos ajustam-se até que o modelo seja capaz de predizer, mais precisamente possível com os testes de hipóteses, a probabilidade de ocorrência de sítios arqueológicos. No final dessa etapa, se faz necessário para a validação do modelo preditivo, realizar a procura dos materiais em campo que poderão também acrescentar ao modelo novas observações.

No último estágio o pesquisador deve analisar os resultados obtidos com a aplicação do modelo preditivo, podendo ser introduzidas novas variáveis para enriquecer e solidificar a integridade do modelo preditivo ajustado.

Em resumo, algumas diretrizes devem ser estabelecidas para a elaboração de um modelo preditivo. Warren (1990) apresenta uma visão geral dos métodos de modelagem preditiva que têm sido populares entre os arqueólogos durante os últimos anos, destacando modelos de probabilidade, nos quais a variável dependente consiste em dois ou mais grupos exaustivos e mutuamente exclusivos, que podem ser codificados como sucesso *versus* fracasso, resposta *versus* não resposta, presença *versus* ausência, etc., com respeito a uma ou mais variáveis independentes. São eles: modelos de decisão por árvore, análise de regressão múltipla e análise de regressão logística. Dando continuidade a este capítulo serão apresentados os dois primeiros métodos de predição. A análise de regressão múltipla é uma técnica ampla, mas de interesse quando a variável resposta (dependente) é contínua. Neste trabalho, a variável de interesse consiste em apenas dois grupos mutuamente exclusivos. Por essa razão, com relação aos modelos de regressão, será abordado apenas o procedimento de regressão logística, que ajusta melhor essas predições. Kvamme (1990), Warren (1990) e Pereira e Itami (1991), entre outros, obtiveram bons resultados com a integração de Sistema de Informação Geográfica com procedimento de analise de regressão logística. Por essa razão, será dado um enfoque especial a este último modelo, apresentando-o em um capítulo separado.

3.1.1. Método de decisão por árvore

Dependendo do problema, o propósito básico de classificação é produzir um classificador acurado ou revelar a estrutura preditiva do problema (Breiman et. al., 1984). Então, o objetivo do método de decisão por árvore é predizer ou apresentar explicações sobre respostas de uma variável dependente, tendo algo em comum com outras técnicas de classificação como Análise Discriminante, a Estatística Não Paramétrica[1] e a Estimação Não Linear.

O método de decisão por árvore tem como analogia uma árvore, na qual a raiz pode ser considerada como o ponto de partida da análise. Esse método é utilizado para

[1] Os parâmetros são quantidades que caracterizam uma distribuição. Na Estatística Não Paramétrica não é necessário fazer suposições sobre a distribuição desses parâmetros da qual tenham sido extraídos os dados para análise, por exemplo, se a distribuição da população é normal.

predizer casos ou classes, a partir de uma variável categórica e dependente e uma ou mais variáveis preditoras.

Na medida em que o conjunto amostral vai sendo comparado, os elementos vão sendo classificados com base no conhecimento existente sobre eles.

Quando uma solução for gerada por decisão de árvore, seu resultado pode ser capaz de predizer novos dados e assim ser considerado um modelo de predição.

Como uma árvore, esse tipo de classificador também tem ramificações, folhas e raiz. A raiz, como dito anteriormente, é o primeiro teste feito no conjunto de dados. As ramificações são os testes subseqüentes ao primeiro e caso a observação esteja na última instância da ramificação, ela será dita como pertencente à classe determinada por esse ramo. Quando a análise determina a classe do registro ou elemento analisado, pode-se afirmar que ele permanece no nó final ou na folha da árvore que compõe os elementos classificados.

A figura 09 mostra a estrutura do método de decisão por árvore.

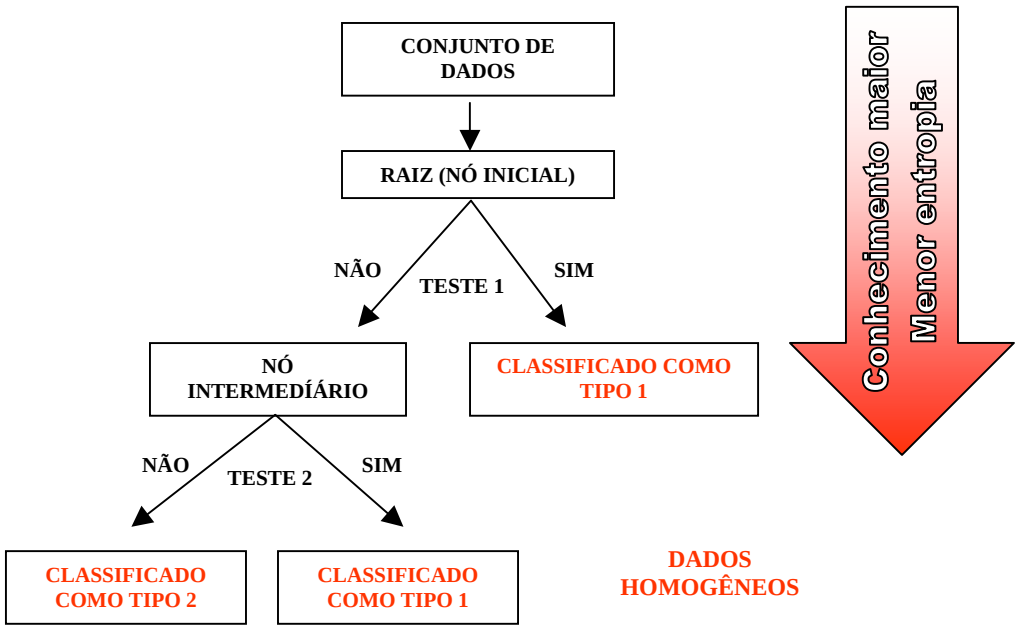


Figura 09: Árvore de classificação

Uma árvore de decisão utiliza uma estratégia de *dividir para encontrar*:

– Um problema complexo é decomposto em sub-problemas mais simples;
– Recursivamente a mesma estratégia é aplicada a cada sub-problema.

A capacidade de discriminação de uma árvore vem da:

– Divisão do espaço definido pelos atributos em sub-espaços;
– Do fato que a cada sub-espaço é associada uma classe;

Como visto acima, os nós da árvore são verificados através de comparações, assim sendo, o problema maior é determinar as regras a serem utilizadas nas comparações, pois a robustez do modelo envolve diretamente o resultado da predição.

Uma estratégia adequada para a tomada de decisão gera um modelo super ajustado, perfazendo uma melhor classificação e assim, os dados ficam próximos da realidade.

Para a geração de uma estratégia "robusta" é necessário realizar uma modelagem que decida a melhor escolha para classificar as variáveis.

3.1.2. Modelagem através de decisão por árvore

A modelagem, neste caso, trata de um processo que determina a seqüência dos nós da árvore. A decisão é obtida de acordo com a prioridade dos nós (testes). Ela permite que o método de decisão por árvore fique mais maleável de forma a aceitar os valores ambíguos ou *outliers* (valores discrepantes). O recurso utilizado para esse tipo de situação é denominado "poda" que, segundo Carvalho (2001), tenta generalizar a árvore eliminando sub-árvores que parecem ser muito específicas.

A seguir é feita uma abordagem do método de decisão por árvore ID3, desenvolvido por Quinlan em 1993, que utiliza o critério de entropia para dividir os nós, ou seja, parte do princípio de que a entropia cresce com a probabilidade associada a um determinado estado.

O conceito de entropia é amplamente utilizado em Física, mais precisamente em Termodinâmica. Em Estatística está relacionado com a quantidade de informação para explicar um determinado evento. Como exemplo, ao jogar uma cartela na Mega Sena a probabilidade de acertar o primeiro jogo é quase nula, de acertar a quina é pequena, mas é maior do que acertar as seis dezenas e de acertar uma quadra é pequena, mas é maior que todas as anteriores. Sendo a probabilidade de acertar a sena quase nula, a entropia pode ser 1 (quantidade de informação nula), para a quina a entropia pode ter valor 0,96 e para a quadra a entropia pode ser 0,92. Nota-se assim, que a entropia pode ser vista como um valor associado à necessidade de informação para a explicação de um evento. Onde há maior

entropia não é necessário uma quantidade grande de informação para explicar um determinado acontecimento.

No caso da classificação pelo método de árvore, quando a entropia é nula, significa que os dados são homogêneos (mesma classe).

No caso do método ID3, dado um determinado nó *x*, o critério de divisão usado é:

$$Entropia(x) = \sum_i \left(- p_i \log_2 p_i\right) \quad (1)$$

onde p_i é a probabilidade da *i-ésima* classe dentro do nó *x*. Então, para expressar a proporção de informação gerada pela divisão, utiliza-se a seguinte razão:

$$\text{Razão Ganho } (x) = \text{ganho}(x)/\text{Informação da divisão}(x) \quad (2)$$

onde o ganho é a diferença de entropia de informação do nó *x* e o denominador refere-se as divisões realizadas para o nó *x*.

$$\text{Ganho } (x) = \text{entropia}(x) - \text{entropia } (x1, x2)$$

$$\text{Entropia}(x1, x2) = - (p1*\log(p1) + p2*\log(p2)) \quad (3)$$

O valor resultante da razão ganho das variáveis preditoras mostra qual variável deve ser testada em ordem de nós, contada da raiz (maior valor da razão ganho) até o nó mais distante (menor valor da razão do ganho).

Demonstra-se a seguir um exemplo retirado de Carvalho (2001). Supondo que uma locadora de carros queira classificar os seus clientes de acordo com aluguel de carros importados ou nacionais utilizando-se as variáveis preditoras: idade e salário. 17 clientes alugaram carros nacionais (N) e 7 alugam carros importados (I).

Para iniciar a classificação é necessário escolher algumas variáveis para serem utilizadas na formação de subgrupos mais homogêneos. No exemplo, o autor escolheu a idade como nó inicial, verificando-se que os clientes tinham entre 18 a 70 anos. Foram produzidos alguns valores de idade: 20, 40, 50 e 60. Com a idade de 20 anos, dividiu-se o

grupo de clientes em dois subgrupos: aqueles que têm mais de 20 anos e aqueles que têm menos de 20 anos. A entropia dos dois subgrupos foi calculada, obtendo-se a entropia total da informação neste primeiro nível da árvore de decisão. Suponha que a entropia tivesse o valor H_{20} = 0,78. Partindo-se do grupo inicial, novamente, dois novos subgrupos foram criados: aqueles com mais de 40 anos e aqueles com menos de 40 anos. Novamente, calculando-se a entropia desses subgrupos obteve-se o valor H_{40} = 0,64.

Repetindo-se o processo para as idades de 50 e 70 anos, os valores da entropia obtidos foram H_{50} = 0,71 e H_{60} = 0,81. Comparando-se os valores calculados das entropias, observou-se que a menor entropia era do subgrupo 40, ou seja, se utilizar uma regra com base na idade menor ou igual que 40 anos, obtêm-se dados homogêneos.

O próximo passo foi definir um novo critério para a separação dos dados remanescentes ao primeiro nó (teste). A variável salário pôde ser aplicada no próximo nó, onde os valores variavam de R$ 2.000,00 a R$ 10.000,00.

Do mesmo modo à primeira aplicação, os grupos foram subdivididos em: R$ 3.000,00; R$ 5.000,00 e R$ 8.000,00 e foram calculadas as entropias associadas, obtendo-se $H_{3.000}$ = 0,61, $H_{5.000}$ = 0,21 e $H_{8.000}$ = 0,13. Então, para o teste do último nó puderam ser utilizados salários maiores e menores que R$ 8.000,000.

A árvore de classificação desse exemplo é mostrada na figura 10.

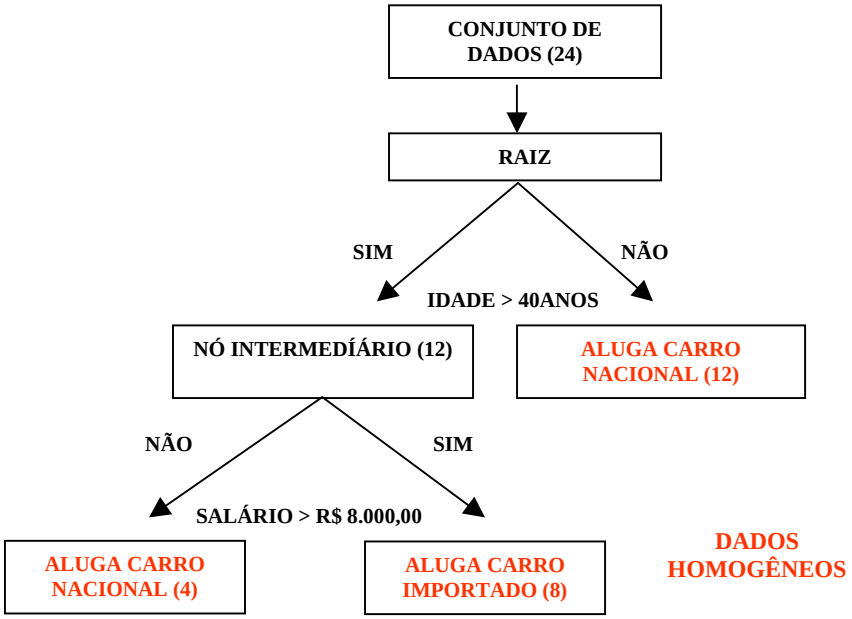


Figura 10: Árvore de Classificação para o exemplo de loja de aluguel

Com base nos dados de treinamento da árvore de decisão, pôde ser gerado um modelo que permitiu classificar um conjunto de dados multivariados; baseado na razão de ganho, uma árvore, contendo testes, permitiu classificar dados localizados na área de confusão entre classes.

Em Arqueologia Limp e Carr (1985) apud Warren (1990) utilizaram essa técnica. Eles selecionaram 13 variáveis ambientais que podiam ter sido importantes para explicar o assentamento pré-histórico no sudeste do Arkansas (EUA), bem como os registros arqueológicos. Cada variável ambiental foi então dicotomizada em alternativa sim-não (decisão favorável *versus* decisão desfavorável) e as dicotomias foram agrupadas em blocos para criar uma série vasta de árvores de decisões hierárquicas. Cada árvore diferia de todas outras em termos de número ou posição de seus blocos. Em seguida, uma amostra de várias estruturas possíveis de decisão por árvore foi aplicada aos dados ambientais na região para identificar "locais viáveis" de povoamento. Essas localidades foram comparadas com as localizações de sítios arqueológicos conhecidos para verificar o potencial preditivo de cada árvore. Uma vantagem desse modelo é a possibilidade da sua estrutura imitar o processo de tomada de decisões de povos pré-históricos, sendo bastante flexível em comparação com métodos tradicionais. Segundo os autores, a desvantagem é que sua implementação é enfadonha e é muito difícil de testar os resultados. Além disso, ele é limitado pelo fato que quando variáveis contínuas são dicotomizadas nas ramificações, o número possível de ramos torna-se infinito.

Também recomenda como último recurso utilizado quando a análise exploratória ou métodos tradicionais não conseguem representar o Mundo Real.

3.2. Análise discriminante pelo método de Fisher

Um outro método de predição utilizando-se a classificação é a análise discriminante, cujo princípio é descrever graficamente (em 3 dimensões ou menos) ou algebricamente as características diferenciais de um conjunto de observações multivariadas, sendo que cada uma dessas observações traz informações de p variáveis $\mathbf{X}$ e estão definidas no espaço p-dimensional R^p.

Essa espécie de classificação permite alocar observações em duas ou mais classes rotuladas, tentando encontrar "discriminantes", cujos valores numéricos fazem com que as populações estejam tão separadas quanto possível (Tachibana, 2000).

Considere *g* populações ou grupos π_1, ..., π_g, onde $g \geq 2$. Supõe-se que a cada população π_j está associada uma função densidade de probabilidade (*fdp*) $f_j(x)$ no espaço R^p, ou seja, se um indivíduo pertence a uma população π_j, tem fdp $f_j(x)$. Então, a meta da análise discriminante é alocar um indivíduo para um dos *g* grupos com base nas observações **x**.

Na teoria apresentada a seguir é dada ênfase no caso de duas populações, ou seja, g = 2, pois é a situação mais comum na Arqueologia, onde uma é a provável área de ocupação pretérita e o seu complementar, a provável área que não foi ocupada.

3.2.1. Função discriminante de Fisher para 2 populações

A idéia da função discriminante de Fisher é transformar a observação multivariada **X** em univariada Y, tal que Y traga informação das populações π_1 e π_2. Se essas populações forem mais distintas possíveis, fica mais fácil afirmar a qual delas pertence as observações; mas nem sempre isso acontece e as populações ocupam algumas áreas em comum no espaço, denominadas "regiões de confusão".

Para resolver esse problema, Fisher, em 1936, sugeriu tomar a combinação linear de **X** para criar Y ($y = \hat{l}' x$), por ser uma função simples de **X** e de fácil tratamento matemático. Tendo μ_{1y} ($E(l'x/\pi_1)$) como a média dos resultados Y, obtida das **X**, cujas observações pertencem a π_1 e μ_{2y} ($E(l'x/\pi_2)$) a média de Y obtida de **X** que pertence a π_2, Fisher selecionou a combinação linear que maximiza o quadrado da distância entre μ_{1y} e μ_{2y} relativa à variabilidade de **X** nas duas populações, dadas pelas matrizes de covariância $\Sigma = E_i[(x-\mu_i)(x-\mu_i)']$, i = 1, 2, considerada igual para duas populações. Nessa matriz, μ_1 e μ_2 são, respectivamente, a média da população de **X** da população π_1 e média de **X** da população π_2.

A distância máxima das duas populações é dada por $(x-\mu_1)'\Sigma^{-1}(x-\mu_2)$. Naturalmente as quantidades populacionais μ_1, μ_2 e Σ raramente são conhecidas e a expressão anterior só poderá ser utilizada se forem estimadas as quantidades populacionais.

Têm-se n_1 observações da variável multivariada **X'** = [x_1, x_2, ..., x_p] de π_1 e n_2 medidas dessa quantidade de π_2. Sejam as seguintes estatísticas relativas as amostras, denotando, respectivamente, a média amostral e variância amostral.

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad \text{e} \quad \mathbf{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \quad , i = 1, 2 \tag{4}$$

A função discriminante de Fisher é construída sem assumir a existência de uma função de probabilidade associada a cada grupo.

Fisher escreveu uma função linear **y=l'x**, que maximiza a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos.

Porém, se as duas populações têm uma matriz de variância e covariância comum, a matriz **S** pode ser substituída pela matriz $\boldsymbol{S}_{pooled.}$ (combinado)

$$\boldsymbol{S}_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2 \tag{5}$$

Para alocar o objeto na população π_1, primeiramente há a necessidade de definir o ponto médio $\hat{m}$ da combinação linear, ou seja $\hat{m}$ = ½$(\bar{x}_1 - \bar{x}_2)$ = ½$(\bar{x}_1 - \bar{x}_2)'\mathrm{S}^{-1}_{\mathrm{pooled}}(\bar{x}_1 - \bar{x}_2)$. Então uma observação x_0 será classificada como pertencente à população π_1 se

$$(\bar{x}_1 - \bar{x}_2)'\mathbf{S}^{-1}_{pooled}\, x_0 \geq \hat{m} \tag{6}$$

e alocada para o grupo π_2 se $(\bar{x}_1 - \bar{x}_2)'\mathbf{S}^{-1}_{pooled}\, x_0 < \hat{m}$

A figura 11 mostra claramente a solução de Fisher para o problema de separação e classificação para p = 2. O conjunto de dados não fica tão discriminado se projetados nos eixos x_1 e x_2, assim sendo, o método rebate os dados numa função linear dos dois eixos. A melhor decisão é a que torna máxima a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos.

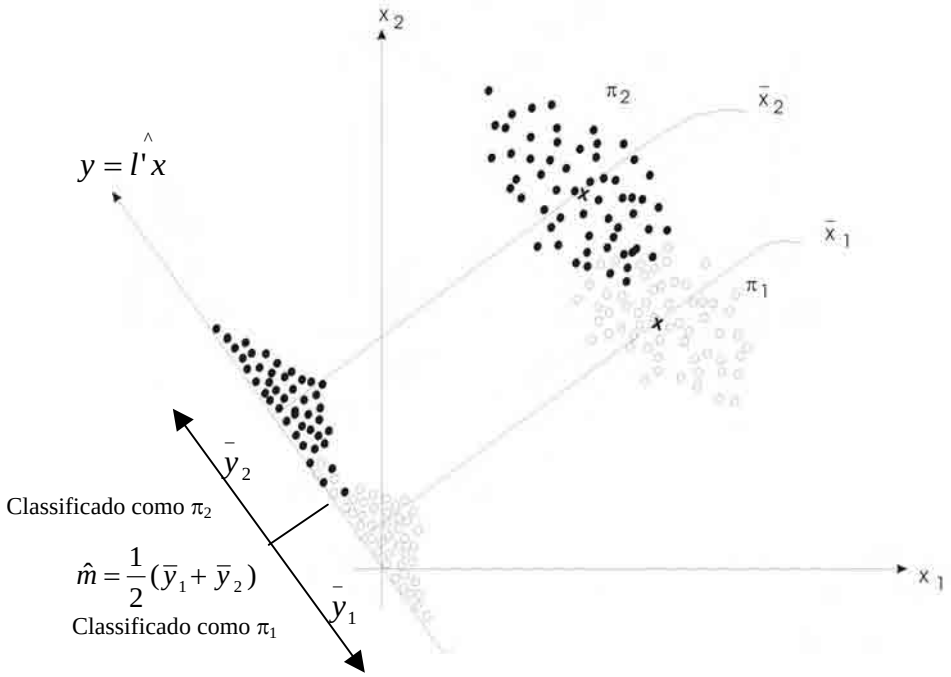


Figura 11: Processo de classificação pelo método de Fisher para 2 populações (Johnson e Wichern, 1992)

Segundo Johnson e Wichern (1992), este tipo de análise só faz sentido se as duas populações realmente tiverem médias diferentes. Suponha que as populações π_1 e π_2 sejam normais multivariadas com uma matriz de covariância comum Σ. Um teste de H_0 : $\mu_1=\mu_2$, contra H_0 : $\mu_1 \neq \mu_2$ é dado pela estatística $\{(n_1+n_2-p-1)/[(n_1+n_2-2)p)]\}\{(n_1n_2)/(n_1+n_2)\}D^2$ que tem distribuição F com $v_1 = p$ e $v_2 = n_1+n_2-p-1$ graus de liberdade e $D^2=(\bar{x}_1-\bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1-\bar{x}_2)$. Se H_0 for rejeitada, pode-se concluir que a separação entre as duas populações π_1 e π_2 é significante, caso contrário significa que as duas populações têm a mesma média e covariância, ou seja, elas formam uma única população.

Um exemplo para esclarecer o método discriminante de Fisher é apresentado por Johnson e Wichern (1992), no qual foi realizado um experimento para detectar portadoras do gene de hemofilia.

Nessa análise foram utilizadas 2 populações, sendo que o primeiro grupo (n_1= 30) era formado por mulheres que não tinham o gene da hemofilia e no segundo (n_2 = 22) eram portadoras do gene de hemofilia.

As variáveis medidas nos grupos foram a atividade AHF {X_1=log(atividade AHF)} e o antígeno AHF {X_2=log(antígeno AHF)}. Suponha que:

$$\bar{x}_1 = \begin{bmatrix} -0{,}0065 \\ -0{,}0390 \end{bmatrix},\ \bar{x}_2 = \begin{bmatrix} -0{,}2483 \\ 0{,}0262 \end{bmatrix} \text{ e } \mathbf{S}^{-1}{}_{pooled} = \begin{bmatrix} 131{,}158 & -90{,}423 \\ -90{,}423 & 108{,}147 \end{bmatrix}$$

sendo a função discriminante de Fisher $y = \hat{l}'x = (\bar{x}_1 - \bar{x}_2)'\mathbf{S}^{-1}_{pooled}x_0 =$

$$\begin{bmatrix} 0{,}2418 & -0{,}0652 \end{bmatrix} \begin{bmatrix} 131{,}158 & -90{,}423 \\ -90{,}423 & 108{,}147 \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 37{,}6x_1 - 28{,}92x_2$$

Assim,

$$\bar{y}_1 = \hat{l}'\bar{x}_1 = \begin{bmatrix} 37{,}6 & -28{,}92 \end{bmatrix} \begin{bmatrix} -0{,}0065 \\ -0{,}0390 \end{bmatrix} = 0{,}88$$

$$\bar{y}_2 = \hat{l}'\bar{x}_2 = \begin{bmatrix} 37{,}6 & -28{,}92 \end{bmatrix} . \begin{bmatrix} -0{,}2483 \\ 0{,}0262 \end{bmatrix} = 10{,}10$$

Com as médias das duas populações projetadas na reta y, determina-se que o ponto médio entre essas médias é

$$\hat{m} = -\frac{1}{2}\left(\bar{y}_1 + \bar{y}_2\right) = \frac{1}{2}\left(0{,}88 - 10{,}10\right) = -4{,}61$$

Com base no ponto médio, as mulheres podem ser classificadas na população π_1 e π_2, aplicando-se apenas uma regra de desigualdade.

3.2.2. Problema de classificação geral.

Quando se utiliza uma reta para classificar os grupos, podem ser gerados erros de determinada grandeza, ou seja, um elemento x pode ser classificado em uma determinada população, mas na realidade ele não pertence a essa população na qual foi designada. Esse tipo de erro pode ocorrer principalmente na região fronteira entre as populações, cujas características dos elementos (observações) sejam muito semelhantes.

Um outro tipo de problema de classificação é o da ordem custo, pois o erro de classificar um objeto π_1 como pertencente a uma classe π_2, talvez seja mais caro que o de classificar um objeto π_2 em uma classe π_1.

Sejam $f_1(x)$ e $f_2(x)$ funções densidades de probabilidade associadas ao vetor *px1* de variáveis aleatórias **X**, respectivamente das populações π_1 e π_2. Um objeto, com medida x associada, deve ser alocado para π_1 ou π_2. Seja Ω o espaço amostral, isto é, todo conjunto de possíveis observações de **x**. Seja R_1 o conjunto de valores de x que são classificados como objetos de π_1 e $R_2 = \Omega - R_1$ os restantes valores de x que são classificados como objetos de π_2.

$$\Omega = R_1 \cup R_2$$

Dessa forma todo o objeto deve ser associado a uma e somente uma das duas populações, os conjuntos R_1 e R_2 são mutuamente excludentes.

A probabilidade condicional, p(2/1), de classificar um objeto como π_2 quando na verdade é de π_1 é:

$$p(2/1) = p(x \in R_2 / \pi_1) = \int_{R_2 = \Omega - R_1} f_1(x)dx \qquad (7)$$

Similarmente, a probabilidade condicional, p(1/2), de classificar um objeto como π_1 quando ele vem de π_2 é

$$p(1/2) = p(x \in R_1 / \pi_2) = \int_{R_1 = \Omega - R_2} f_2(x)dx \qquad (8)$$

A integral da expressão 7 representa a área formada pela função densidade $f_1(x)$ sobre a região R_2. Similarmente, a integral $p(1/2)$, representa a área formada por $f_2(x)$ sobre a região R_1, como mostra a figura 12 para o caso univariado.

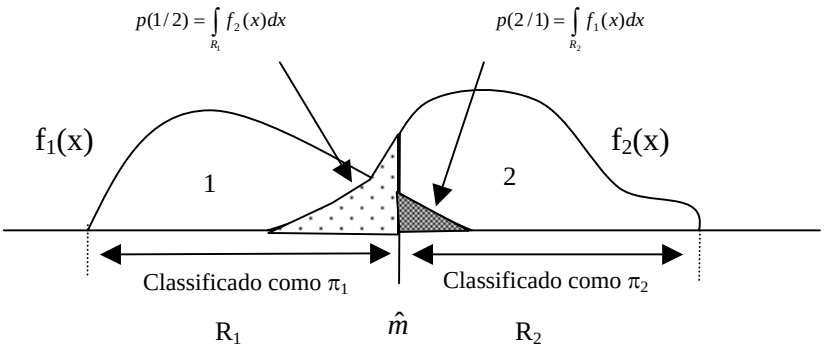


Figura 12: Probabilidades de classificação errada para regiões de classificação quando p=1 (Johnson e Wichern, 1992)

Seja P_1 a probabilidade a priori de π_1 e P_2 a probabilidade a priori de π_2 , então P1 + P2 = 1.

A probabilidade de classificar corretamente, ou incorretamente os objetos pode ser derivada como o produto das probabilidades a priori das probabilidades condicionais.

1 - P(classificada corretamente como π_1) = P(observação vem de π_1 e é classificada corretamente como π_1) = $P(x \in R_1 / \pi_1).P(\pi_1) = P(1/1).P1$

2 – P(classificada incorretamente como π_1) = P(observação vem de π_2 e é classificada incorretamente como π_1) = $P(x \in R_1 / \pi_2).P(\pi_2) = P(1/2).P2$

3 – P(classificada corretamente como π_2) = P(observação vem de π_2 e é classificada corretamente como π_2) = $P(x \in R_2 / \pi_2).P(\pi_2) = P(2/2).P2$

4 – P(classificada incorretamente como π_2) = P(observação vem de π_1 e é classificada incorretamente como π_2) = $P(x \in R_2 / \pi_1).P(\pi_1) = P(2/1).P1$

Mesmo uma probabilidade aparentemente pequena como 0,02 = P(2/1), a probabilidade pode tornar-se relativamente grande se o custo de fazer uma atribuição incorreta para π_2 for extremamente alto.

O custo da classificação errada pode ser definido em uma matriz de custos.

Tabela 01: Custo da classificação errada

	Classificado como		
		π_1	π_2
População Verdadeira	π_1	0	C(2/1)
	π_2	C(1/2)	0

Os custos são:

1. Zero para classificação correta.
2. C(1/2): Observação de π_2 incorretamente classificada como π_1, e C(2/1): Observação de π_1 incorretamente classificada como π_2.

O valor esperado do custo de uma classificação incorreta {E(CM)} é o produto da diagonal secundária por suas probabilidades de ocorrência,

$$E(CM) = C(2/1)\ P(2/1)\ P1 + C(1/2)\ P(1/2)\ P2 \qquad (9)$$

devendo ressaltar que o E(CM) deve ser o menor possível.

As explicações demonstradas, para problemas de erros de classificação do método de Fisher, foram tratadas como dados univariados somente e duas populações, para um fácil entendimento. Retornando para dados multivariados, segundo Johnson e Wichern (1992), Fisher propôs uma extensão do seu método discriminante para várias populações. A extensão assume que as matrizes *pxp* de variâncias e covariâncias das *g* populações são iguais e de posto completo. Sejam $\overline{\mu}$, a média dos grupos combinados e B a soma dos produtos cruzados entre grupos, em que:

$$\bar{\mu} = \frac{1}{g}\sum_{i=1}^{g}\mu_i \quad \text{e} \quad \text{B}=\sum_{i=1}^{g}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'$$

cujos vetores amostrais e as matrizes amostrais são respectivamente

$$\bar{x} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{g} n_i} \quad \text{e} \quad \hat{B} = \sum_{i=1}^{g}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{10}$$

A matriz de variação dentro dos grupos é dada por $W = \sum_{i=1}^{g}(n_1 - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$.

Sejam $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_\Delta > 0$, $\Delta \leq \min(g-1, p)$ autovalores não nulos de $W^{-1}\hat{B}$ e $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_s$ os correspondentes autovetores. O vetor de coeficientes que maximiza a razão entre a soma de quadrados entre grupos e a soma de quadrados dentro dos grupos, ou seja, que maximiza

$$\frac{\hat{l}'\hat{B}\hat{l}}{\hat{l}'W\hat{l}} = \frac{\hat{l}'\left(\sum_{i=1}^{g}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'\right)\hat{l}}{\hat{l}'\left(\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{x}_{ij} - \bar{x}_i)(\bar{x}_{ij} - \bar{x}_i)'\right)\hat{l}} \tag{11}$$

$$\hat{B} = \sum_{i=1}^{g}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \quad \text{, onde} \tag{12}$$

é dado por $\hat{e}_1 = \hat{l}_1$. A combinação linear $\hat{l}_1'$ é chamada primeiro discriminante amostral. O segundo discriminante amostral é dado por $\hat{l}_2'$, onde $\hat{l}_2 = \hat{e}_2$. O *k-ésimo* discriminante amostral é dado por $\hat{l}_k'$, onde $\hat{l}_k = e_k$, $k \leq s$. Desta forma, baseado nos primeiros $r \leq s$ discriminantes amostrais, o objeto com variáveis medidas x_0 é alocado à população π_k se

$$\sum_{j=1}^{r}[\hat{l}_j'(x_0 - \bar{x}_k)]^2 \leq \sum_{j=1}^{r}[\hat{l}_j'(x_0 - \bar{x}_i)]^2 \text{ para todo } i \leq k \tag{13}$$

Os métodos demonstrados acima foram amplamente utilizados por arqueólogos no passado para predizer as prováveis áreas de ocupação pretérita, porém como a procura de vestígios arqueológicos trata-se de uma variável resposta do tipo dicotômica, o capítulo seguinte abordará sobre o método de Regressão Logística.

4. REGRESSÃO LOGÍSTICA

No capítulo anterior foram vistos alguns métodos de predições utilizados para determinar a probabilidade de ocorrência de vestígios arqueológicos. Porém, esses métodos são classificadores, ou seja, separam, a partir características mais semelhantes, em determinados grupos, podendo fornecer, após uma posterior aplicação, um modelo de predição para novos dados independentes baseado nessa separação de populações.

Nesse capítulo será apresentado o método de Regressão logística, o qual permite realizar uma modelagem preditiva, usando um modelo de dados binários que calcula a probabilidade referente ao sucesso de um determinado evento.

A regressão logística é um modelo probabilístico que descreve a relação entre uma variável resposta e uma ou mais variáveis explicativas, apresentando a resposta de maneira sucinta, geralmente como um número ou uma série de números.

No campo da Estatística são definidos os significados dos números, ou seja, a interpretação dos relacionamentos entre uma variável resposta, geralmente binária ou dicotômica e uma ou mais variáveis explicativas. Essa relação determina a probabilidade de ocorrência de um evento em presença de um conjunto de variáveis independentes e explicativas, formando um modelo preditivo indutivo, no qual uma função é explicada por observações empíricas.

Berkson apude Warren (1990), o modelo de regressão logística começou a ser difundido nos últimos anos pela alta capacidade de processamento dos computadores e pelo desenvolvimento de pacotes estatísticos.

Essas aplicações têm sido importantes em áreas como Economia, Mineração, Transportes, Sensoriamento Remoto, Medicina e nas Ciências Sociais.

Nesse capítulo está apresentada a regressão logística, partindo do ponto de vista de regressão linear, destacando-se os pontos comuns e as diferenças entre esses modelos, pois desse modo facilitando a compreensão.

4.1. Modelo de regressão logística

Segundo Hosmer e Lemeshow (1989), em problemas de regressão, a quantidade chave é o valor médio da variável resposta dado o valor da variável independente, que é chamada de média condicional e expressa pela esperança de Y, dado um valor x, E(Y/X=x). É freqüente o caso em que a variável resultante é discreta, podendo assumir dois ou mais valores possíveis.

Em regressão linear assume-se que esta média possa ser expressa como uma equação linear em x ou alguma transformação de X ou em Y, tal como:

$$E(Y/x) = \beta_0 + \beta_1 x \qquad (14)$$

Assim sendo, é possível que a média assuma qualquer valor quando *x* varia entre $-\infty$ e $+\infty$.

Hosmer & Lemeshow (1989) exemplificam um caso, no qual observaram-se duas variáveis: a idade das pessoas e se elas tinham problemas cardíacos (CHD). O gráfico dessas duas variáveis mostra que todos os pontos caíram em uma das duas linhas paralelas, representando a ausência de CHD (y = 0) e a presença de CHD (y = 1). Existe alguma tendência para os indivíduos com nenhuma evidência de CHD ser mais jovem do que aquelas com evidência de CHD. Enquanto este gráfico descreve a natureza dicotômica da variável resultante, claramente ele não fornece um aspecto claro da natureza da relação entre CHD e idade.

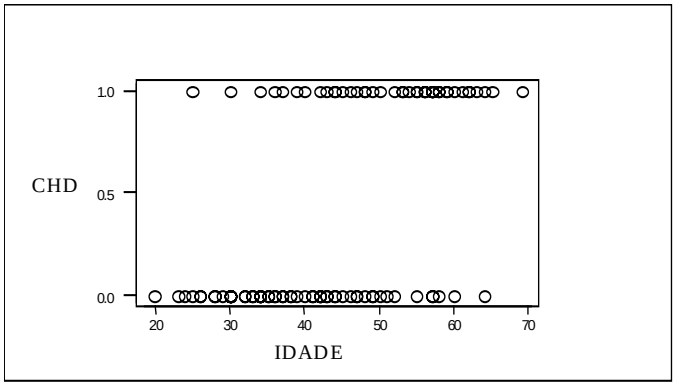


Figura 13: Gráfico de problema cardíaco e idade (Hosmer & Lemeshow, 1989)

A grande variabilidade em CHD e idade dificulta a interpretação dos dados. Para remoção dessa variação, mantendo-se a estrutura da relação entre a resposta e a variável independente, a variável independente, apresentada na Tabela 02, foi dividida em várias faixas ou classes. Em seguida calculou-se a média condicional da variável resposta em cada grupo.

Tabela 02: Freqüência de problemas cardíacos por faixa etária

CHD				
Grupo de Idade	**n**	**Ausente**	**Presente**	**Média(proporção)**
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
TOTAL	100	57	43	0.43

Fonte: Hosmer e Lemeshow, 1989

Examinando a Tabela 02, um aspecto mais claro da relação começa a surgir. Parece que, com o aumento da idade, a proporção de indivíduos com problemas cardíacos (CHD) aumenta. A figura 12 apresenta um gráfico da proporção de indivíduos com problemas cardíacos

contra o ponto médio de cada grupo de idade. Esse gráfico fornece uma melhor compreensão da relação entre CHD e idade.

Quando a variável resposta é dicotômica[1], sua média condicional deve ser maior ou igual a zero e menor ou igual a 1 [$0 \leq E(Y/x) \leq 1$], aproximando-se de 0 e de 1 gradualmente (forma de "S") e cujo gráfico se parece com uma distribuição acumulada, isso fica bem evidenciado na Figura 14.

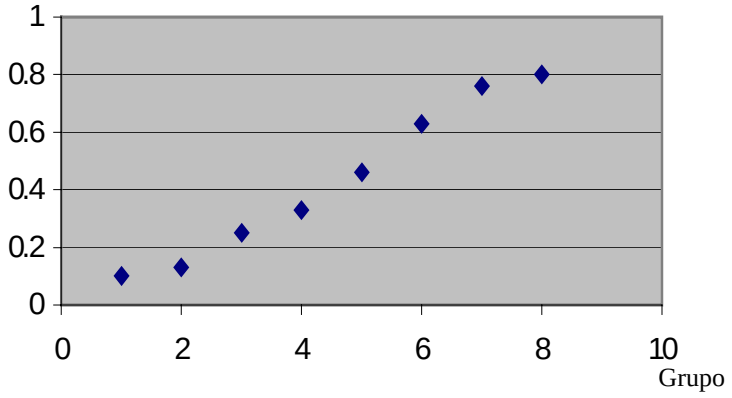


Figura 14: Proporção (p) de pessoas com problemas cardíacos nas diferentes faixas etárias.

Portanto, no modelo de regressão linear assume-se que a variável resposta é contínua, enquanto que no modelo de regressão logística a variável resposta é discreta: binária ou dicotômica.

Assim sendo, quando o vetor da média condicional E(Y/X) pode assumir qualquer valor quando X varia entre $-\infty$ e $+\infty$, os parâmetros do modelo podem ser estimados usando-se o método dos Mínimos Quadrados (MMQ), pois se está ajustando um modelo linear. Porém, quando o vetor da média condicional apresentar a forma de uma distribuição acumulada, como no caso da variável dicotômica a estimação dos parâmetros da função é dada pela máxima verossimilhança (não linear).

Segundo os estudos de Cox apud Hosmer e Lemeshow (1989), muitas das funções distribuições têm sido propostas, porém a função ideal para o caso da variável resposta

[1] Variável que pode assumir somente dois valores possíveis (um ou zero, sim ou não, ...)

ser dicotômica, é a função logito (*logit*), pois é extremamente flexível e fácil de ser usada e interpretada.

Então, a forma do modelo de regressão logística é baseada no modelo *logit* e para dados univariados é dada como:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{15}$$

Sendo o modelo acima não linear, aplicou-se uma transformação denominada *g(x)* para obter algumas propriedades "desejáveis" do modelo de regressão linear, como:

- A soma dos valores $\hat{y}$ é igual a soma dos valores de y $\sum_{i-1}^{n} y_i = \sum_{i-1}^{n} \hat{y}_i$;

- A soma dos erros é igual a zero, $\sum_{i-1}^{n} e_i = 0$;

- A soma das observações multiplicada pelo erro é igual a 0, $\sum_{i=1}^{n} x_i e_i = 0$

- A reta de mínimos quadrados (MQ) passa pelo ponto ($\bar{x}, \bar{y}$).

Assim, a transformação *g(x)* torna o modelo *logit* linear em seus parâmetros contínuos e faz com que assuma valores entre -∞ e +∞, dependendo do limite de x:

$$g(x) = ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x \tag{16}$$

No modelo de regressão linear existe um ε que expressa um desvio da observação em relação à média condicional ($y = E(Y/x) + \varepsilon$) , sendo que, comumente, supõe-se que ε segue uma distribuição normal com média zero e alguma variância constante.

Já no caso de variáveis de resposta dicotômica, se y = 1, então $\varepsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$ e se y = 0, então $\varepsilon = - \pi(x)$ com probabilidade $1 - \pi(x)$. Assim, a distribuição

condicional da variável resposta segue uma distribuição binomial com probabilidade dada pela média condicional.

Em resumo, quando a variável resposta é dicotômica:

- a média condicional da equação de regressão deve ser formulada para estar entre zero e 1. Tem-se estabelecido que o modelo de regressão logística $\pi(x)$ satisfaz essa restrição;
- a distribuição binomial, não a normal, descreve a distribuição dos erros e será a distribuição estatística na qual a análise é baseada.

Assim, o princípio que orienta uma análise usando regressão linear poderá orientar na regressão logística. Segundo Homer e Lemeshow (1989), na regressão linear o método mais usado para estimação dos parâmetros β é o dos mínimos quadrados, no qual são determinados valores de β que minimizam a soma dos quadrados de desvios de valores observados de Y dos valores preditos, baseados no modelo.

Quando o MMQ é utilizado em modelo com resultado dicotômico, os estimadores não apresentam as propriedades estatísticas desejáveis.

Para solucionar o problema é utilizado o método da máxima verossimilhança, que produz valores para os parâmetros desconhecidos que maximizam a probabilidade de obtenção dos conjuntos de dados observados.

A função de verossimilhança expressa a probabilidade dos dados observados como uma função de parâmetros desconhecidos e seus estimadores resultantes são aqueles que mais se aproximam do resultado.

Para Meyer (1978), o método de máxima verossimilhança (MV) conduz a estimativas razoáveis para os dados dicotômicos.

4.1.1. Estimativas de Máxima Verossimilhança

Segundo Meyer (1978), a definição da estimativa de máxima verossimilhança de β, isto é, $\hat{\beta}$, baseada em uma amostra aleatória $x_1,...,x_n$ é aquele valor de β que torna máxima $L(x_1,...,x_n; \beta)$, considerada como uma função de β para uma dada amostra $x_1,...,x_n$, e onde L é definida pela equação (15). Essa estimativa é, geralmente, referida como a estimativa de MV.

Assim, seja $x_1,...,x_n$ uma amostra aleatória de X (valores amostrais). A função de verossimilhança L é definida como a seguinte função da amostra e de β:

$$L(X_1,...,X_n; \beta) = f(X_1; \beta)\, f(X_2; \beta) \ldots f(X_n; \beta). \qquad (17)$$

O modelo (16) representa as probabilidades $P[X_1= x_1,...,x_n]$ se X for discreta, enquanto se X for contínua, $L(x_1,...,x_n; \beta)$ representará a função densidade de probabilidade conjunta de $(x_1,...,x_n)$.

A questão primordial, sendo β desconhecido, é:

" Para qual valor de β $L(x_1,...,x_n; \beta)$ será máxima? "

A resposta está baseada no valor do parâmetro que torne o mais provável possível, um evento que já ocorreu, ocorra novamente, como estabelece a definição da estimativa de MV.

Para pares (x_i, y_i) onde $y_i = 1$, a contribuição para a função de verossimilhança é $\pi(x_i)$ e para aqueles onde $y_i = 0$ a contribuição para a função de verossimilhança é $1 - \pi(x_i)$, sendo que a quantidade $\pi(x_i)$ denota o valor de $\pi(x)$ calculado em x_i. Uma forma conveniente de expressar a contribuição da função de verossimilhança para o par é dada pelo termo:

$$\zeta(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \qquad (18)$$

Quando as observações são assumidas independentes, a função de verossimilhança é obtida pelo produto dos termos dados na equação anterior, o produtório oferece a densidade conjunta de X_1, X_2, ..., X_n:

$$L(\beta) = \prod_{i=1}^{n} \zeta(X_i) \tag{19}$$

Para facilitar o tratamento matemático da expressão (19), é aplicada uma transformação logaritmica que transformará o produtório em somatório.

$$L(\beta) = ln[\, l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i)\ln[1 - \pi(x_i)]\} \tag{20}$$

Para determinar os valores de β que maximizam $L(\beta)$, deriva-se a função (20) em relação aos elementos do vetor β, que por sua vez são funções dos logaritmos presentes na equação. A seguir demonstra-se como são descritas as equações de verossimilhança, sabendo-se que (para o caso uma variável explicativa):

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Substituindo os valores de $\pi(x)$ em (20) e utilizando-se as propriedades dos logaritmos neperianos, tem-se:

$$L(\beta) = \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right) + (1 - y_i)\ln\left[1 - \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right)\right]$$

$$\sum_{i=1}^{n} \left\{ y_i \left[\ln(e^{\beta_0 + \beta_1 x_i}) - \ln(1 + e^{\beta_0 + \beta_1 x_i})\right] + \left[-\ln(1 + e^{\beta_0 + \beta_1 x_i}) + y_i \ln(1 + e^{\beta_0 + \beta_1 x_i})\right]\right\}$$

$$= \sum_{i=1}^{n} y_i \ln(e^{\beta_0+\beta_1 x_i}) - \ln(1+e^{\beta_0+\beta_1 x_i}) \tag{21}$$

Para determinar o valor de *L(β)*, diferencia-se *L(β)* com respeito à β_0 e β_1 e faz-se os resultados das expressões iguais a zero (exemplo: caso unidimensional β_0 e β_1):

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} y_i \ln(e^{\beta_0+\beta_1 x_i}) - \ln(1+e^{\beta_0+\beta x_i}) =$$

$$= \sum_{i=1}^{n} y_i - \left[\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}} \right) \left(\frac{\partial}{\partial \beta_0} [1+e^{\beta_0+\beta_1 x_i}] \right) \right] =$$

$$= \sum_{i=1}^{n} y_i - \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) =$$

$$= \sum_{i=1}^{n} y_i - \pi(x_i) \tag{22}$$

Derivando em relação a β_1

$$= \frac{\partial}{\partial \beta_1} \left[\sum_{i=1}^{n} [y_i \ln(e^{\beta_0+\beta_1 x_i}) - \ln(1+e^{\beta_0+\beta_1 x_i})] \right] =$$

$$= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i (\beta_0 + \beta_1 x_i) - \ln(1+e^{\beta_0+\beta_1 x_i})] =$$

$$= \sum_{i=1}^{n} x_i y_i - \left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}} \frac{\partial}{\partial \beta_1} e^{\beta_0+\beta_1 x_i} \right) =$$

$$= \sum_{i=1}^{n} x_i \left[y_i - \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) \right] =$$

$$= \sum_{i=1}^{n} x \ y_i - \left(\frac{x e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) =$$

$$= \sum_{i=1}^{n} x_i [y_i - \pi(x_i)] \tag{23}$$

O valor de β, dado pela solução das equações de verossimilhança, é denominado estimativa de máxima verossimilhança e será denotado como $\hat{\beta}$. Em geral, o uso do símbolo ^ denotará o estimador de máxima verossimilhança (EMV).

As propriedades mais importantes do método máxima verossimilhança são:

- A estimativa de MV pode ser tendenciosa, e com freqüência, tal tendenciosidade pode ser eliminada pela multiplicação por uma constante apropriada;
- Sob condições bastante gerais, as estimativas de MV são coerentes. Isto é, se o tamanho da amostra sobre a qual essas estimativas são calculadas for grande, a estimativa de MV será "próxima" do valor do parâmetro a ser estimado. As estimativas de MV possuem outra propriedade de "grandes amostras" muito importante; isto é, elas são assintóticas;
- As estimativas de MV apresentam a seguinte propriedade de invariância muito importante: Se $\hat{\theta}$ é uma estimativa de MV de θ, então g($\hat{\theta}$) também é uma estimativa de MV de uma função monótona contínua g($\hat{\theta}$). Explicando: A estimativa de g($\hat{\theta}$), onde g é uma função (monótona) de θ é também aquela mesma função da estimativa de MV de θ. Lembrando-se que essa propriedade não é apresentada por estimativas não tendenciosas (Meyer, 1978).

4.1.2. Teste de significância dos coeficientes

Com os coeficientes estimados, deve-se testar a significância das variáveis decorrentes no modelo. Nesse processo está envolvido o teste de hipóteses estatísticas, o qual determina se as variáveis independentes no modelo estão "significantemente" relacionadas com a variável resposta.

Para Hosmer e Lemeshow (1989), uma aproximação para testar a significância do coeficiente de uma variável em qualquer modelo relaciona-se com a seguinte questão: o modelo que inclui a variável em questão diz mais sobre a variável resultante ou resposta, do que o modelo que não inclui a variável?

No processo comparam-se os valores observados da variável resposta com aqueles preditos, por cada um dos dois modelos; o primeiro com a variável presente e o segundo sem essa variável. A função matemática usada para comparar os valores observados e preditos depende do problema em particular.

A comparação entre os valores preditos e observados, usando a função de verossimilhança, é baseada na seguinte expressão.

$$D = -2\ln\left[\frac{\text{Verossimilhança do modelo atual}}{\text{Verossimilhança do modelo saturado}}\right] \tag{24}$$

Sendo denominada como razão de verossimilhança, aplicada em testes de hipóteses devido à sua distribuição ser conhecida. Na equação é aplicado “menos duas vezes seu log”, pois sua distribuição equivale ao *Qui-Quadrado*.

Utilizando as equações (07) e (12) têm-se:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \tag{25}$$

Essa estatística é chamada de *deviance* e avalia o valor ajustado na regressão logística, desempenhando o mesmo papel que a soma de quadrados residuais (SSE – *Sum Square Error*) tem na regressão linear.

Para estimar a significância de uma variável independente, comparam-se o valor de D com e sem a variável independente na equação.

A alteração em *D*, devido a inclusão da variável independente no modelo, é obtido como:

G = *D*(para o modelo sem a variável) – *D*(para o modelo com a variável)

Esta estatística desempenha o mesmo papel na regressão logística, como faz o numerador do teste F na regressão linear, porque a verossimilhança do modelo saturado é comum para ambos os valores de *D* sendo eliminado no cálculo de *G*. Assim, *G* pode ser expresso como:

$$G\text{=-2ln}\left[\frac{\text{(verossimlhanças sem a variável)}}{\text{(verossimlhança com a variável)}}\right] \tag{26}$$

Para o caso específico de uma única variável independente, pode-se mostrar quando esta variável não está no modelo, o estimador de MV de β_0 é $\ln(n_1/n_0)$ onde $n_1 = \sum y_i$ e $n_0 = \sum(1-y_i)$ e que o valor predito é constante, n_1/n.

Neste caso o valor de *G* é:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{\,y_i}\left(1-\hat{\pi}_i\right)^{(1-y_i)}}\right]$$

ou

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i\ln(\hat{\pi}_i)+(1-y_i)\ln(1-\hat{\pi}_i)\right]-\left[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)\right]\right\} \tag{27}$$

Sob a hipótese de que β_1 é igual a zero, a estatística *G* terá uma distribuição qui-quadrado (χ^2) com 1 grau de liberdade. Suposições adicionais matemáticas são necessárias,

mas para o caso acima elas são não restritivas e implica em se ter um tamanho amostral (*n*) suficientemente grande.

Será utilizado o símbolo $\chi^2(v)$ para denotar uma variável aleatória qui-quadrado com *v* graus de liberdade. Olhando para a tabela qui-quadrado, associa-se o valor p e, portanto é possível obter uma declaração da evidência estatística para esta variável.

Segundo Hosmer e Lemeshow (1989), os cálculos do logaritmo de verossimilhança e o teste da razão de verossimilhanças são aspectos característicos de qualquer pacote de regressão logística. Isto torna possível verificar a significância da adição de novos termos no modelo como um assunto de rotina. No caso de uma única variável independente, recomenda-se ajustar primeiro um modelo contendo apenas o termo constante. Em seguida ajustar um modelo contendo a variável independente, bem como a constante. Estes dados originam um novo log de verossimilhança. O teste da razão de verossimilhanças é obtido multiplicando-se a diferença destes dois valores por -2. Este resultado, bem como o p-valor associado à distribuição qui-quadrado, podem ser obtidos na maioria dos pacotes estatísticos.

Para testar o ajuste dos coeficientes também podem ser utilizados outros métodos estatísticos semelhantes ao anterior, como: Teste de Wald e Teste de Escore.

O Teste de Wald é obtido comparando-se o EMV do parâmetro de inclinação, $\hat{\beta}_1$, com uma estimativa do seu erro padrão ($\hat{S}E$ – *Standard Error*). O resultado da razão, sob a hipótese básica $\beta_1 = 0$, terá uma distribuição normal padrão.

$$W = \frac{\hat{\beta}_1}{\hat{SE}\left(\hat{\beta}_1\right)} \tag{28}$$

O *p*-valor bicaudal é P (| z | > *W*), onde z denota uma variável aleatória seguindo uma distribuição normal padrão. Porém, o teste de Wald em alguns casos, rejeita coeficientes significativos.

Pesquisadores examinando a eficiência do Teste de Wald encontraram que, às vezes, este teste tem se procedido de maneira anormal, rejeitando um o coeficiente quando este

era significativo. Eles recomendam que o Teste da razão de verossimilhanças seja usado. Também foi notada a adequação de inferências na regressão logística baseada na estatística de Wald. Ambos os testes, da razão de verossimilhança, *G*, e o teste de Wald, *W*, requerem o cálculo dos EMV para β_1. Para uma única variável esta não é uma tarefa computacional difícil ou onerosa, porém, para conjuntos de dados grandes com muitas variáveis, o cálculo iterativo necessário para obter a estimativa de máxima verossimilhança pode ser trabalhoso.

Um teste para a significância de uma variável que não requer estes cálculos é o Teste de Escore. Os proponentes do teste de Escore citam esta redução de esforço computacional como a sua maior vantagem. Entretanto o uso deste teste é limitado pelo fato de que ele não pode ser obtido facilmente em alguns pacotes. O teste de Escore é baseado na teoria de distribuição de derivadas do log da verossimilhança.

No caso univariado este teste é baseado na distribuição condicional da derivada da equação (22), dada a derivada da equação (21). Neste caso, pode-se escrever uma expressão para o teste de Escore. O teste usa o valor da equação (22), calculado usando-se $\beta_0 = \ln(n_1 / n_0)$ e $\beta_1 = 0$. Como já foi visto, sob os valores desse parâmetros, $\hat{\pi} = n_1 / n = \bar{y}$. Então, o lado esquerdo da equação (22) torna-se $\sum x_i (y_i - \bar{y})$. Pode ser mostrado que a variância estimada é $\bar{y}(1-\bar{y})\sum (x_i - \bar{x})^2$. A estatística do teste para o teste de Escore (ST) é

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n} (x_i - \bar{x})^2}}. \tag{29}$$

e o *p*-valor bicaudal é P (| z | > ST) < valor limite *p* .

O teste rejeita a hipótese de que o coeficiente β não é significativo se $z_{\alpha/2}$, em valor absoluto, for maior que o valor *W* que foi observado. Nesse caso esse coeficiente β é significativo e a variável correspondente X é importante no modelo.

Em resumo, o método para testar a significância do coeficiente de uma variável na regressão logística segue uma aproximação parecida com aquela usada na regressão linear, mas usa a função de verossimilhança para a variável resposta dicotômica.

5. MATERIAL E MÉTODO

A modelagem preditiva com SIG é uma poderosa ferramenta para auxiliar uma das etapas de um Projeto de Salvamento Arqueológico: a prospecção. Nessa etapa são verificadas as ocorrências e vestígios de antigas civilizações em toda área de estudo, que constitui a vistoria *in loco*, utilizando a percepção visual e o conhecimento do pesquisador.

Warren (1990) diz que a predição tem um papel importante no método científico, pois quando um cientista descobre um padrão num conjunto de observações e consegue desenvolver uma hipótese, útil para explicar esse padrão, ela tem implicações preditivas para as observações futuras. As implicações podem ser testadas com dados novos ou independentes e se os dados novos estiverem de acordo com as predições, esse teste dará subsídio para a validação da hipótese. No método científico a predição é um mecanismo para testar as explicações.

A seguir são apresentados o material e o método utilizado na modelagem preditiva na Arqueologia. O foco está no modelo espacial de localização de prováveis áreas de ocupações arqueológicas (Regressão Logística), analisadas com o auxílio das ferramentas de Geoprocessamento, compreendendo desde a abstração do Universo Arqueológico até a aplicação do modelo de Regressão Logística nos modelos espaciais.

5.1. Material

O presente trabalho utilizou dados do Projeto de Salvamento Arqueológico da Hidrelétrica da cidade de Porto Primavera - SP/MS, (Figura 15), situada no extremo oeste do Estado de São Paulo, cuja área é de aproximadamente 144 km^2 (12 x 12 km).

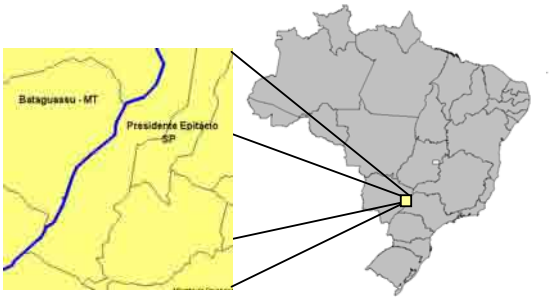


Figura 15: Área de localização

A escolha da área de pesquisa

não foi casual, tendo sido influenciada por alguns fatores, dentre os quais podem ser citados:

- presença ampla de material cartográfico,
- fácil acesso às informações de instituições, Prefeituras e empresas;
- quantidade razoável de ocorrências arqueológicas e
- proximidade com o campus da Unesp de Presidente Prudente.

No Projeto de Salvamento Arqueológico de Porto Primavera, a Companhia de Energia do Estado de São Paulo (CESP) firmou um contrato com a Universidade Estadual Paulista (Unesp), para realizar o Salvamento Arqueológico (Arqueologia de Contrato) na área a ser inundada pela UHE "Sérgio Motta" (Rio Paraná).

Como as informações levantadas referiam-se a uma área de inundação, as mesmas estavam restritas às cotas altimétricas 240 a 270 metros (mínima e máxima), recobrindo uma vasta área no Estado de Mato Grosso do Sul e uma estreita faixa no Estado de São Paulo, devido à conformidade do terreno, por esse fato, foi determinada a aplicação do modelo preditivo somente no lado do Estado de Mato Grosso do Sul.

O material cartográfico utilizado para a elaboração da base de dados digital constituiu-se de (*datum* Córrego Alegre):

- 30 cartas topográficas na escala 1:10.000, baseadas em levantamento aerofotogramétrico de 1965;
- Carta geológica (1979), na escala 1:200.000;
- Carta geomorfológica (1979), na escala 1:200.000;
- Carta pedológica (1979), na escala 1:200.000.

Esses dados foram inseridos em um Banco de Dados Geográficos (DBG), com ferramentas do Arc/Info da *Environmental System Research Institute*, com plataforma o AIX (UNIX) para *Workstations* (IBM *RISC*).

5.2. Modelagem dos dados arqueológicos

A modelagem de dados em um Sistema de Informação Geográfica trata da compreensão de fenômenos geográficos e comportamentais de um determinado universo, requerendo do analista a habilidade de abstração do mundo real para o universo conceitual e posteriormente, para os universos de representação e implementação.

Câmara (1996), diz que um modelo de dados é um conjunto de ferramentas conceituais utilizado para estruturar dados num sistema computacional, que é um aspecto fundamental no projeto de um SIG. O modelo descreve como a realidade geográfica será representada no computador. Nenhuma outra decisão limita tanto a abrangência e o crescimento futuro do sistema quanto à escolha do modelo de dados. Uma modelagem bem elaborada conduz para a otimização de todo o processamento do sistema.

Visto sob essa perspectiva, foi necessário conhecer o Universo Arqueológico através de leituras, estudos e entrevistas com arqueólogos, pois necessitou-se entender o processo de um Projeto de Salvamento Arqueológico, bem como as etapas inerentes ao mesmo, ou sejam a prospecção e escavação.

5.2.1. Problemática

Como a modelagem trata da simplificação do Mundo Real, procurou-se realizar um entendimento das necessidades dos arqueólogos, visando abstrair os conceitos que orientam os arqueólogos nas prospecções realizadas em campo; então, primeiramente, foi determinado o problema que as ferramentas de Geoprocessamento poderiam resolver ou fornecer alternativas de solução:

"Quais os locais mais propícios para encontrar vestígios arqueológicos sem realizar o levantamento in loco em toda área de inundação do lago da UHE?".

Para responder essa pergunta, foram analisadas várias referências bibliográficas, bem como, verificadas as características das populações que habitavam a área de pesquisa.

Segundo Thomaz (2002), a região onde está inserido o módulo de estudo, assim como outras do território nacional, foram intensamente ocupadas, no passado, por grupos de caçadores-coletores e agricultores, principalmente da denominada Tradição Tupiguarani, fato esse constatado por meio de evidências arqueológicas em amplos programas de pesquisa, desenvolvidos pelas instituições e afins. Esses dados são complementados por informações referentes à cultura indígena, transmitida pelos cronistas dos séculos XVIII e XIX.

Assim sendo, a modelagem preditiva aplicada à Arqueologia deve se basear em uma representação adequada da realidade analisada, caso contrário são produzidos modelos inapropriados com erros inerentes à representação utilizada.

5.2.2. Modelo Estático

Para representar a realidade decidiu-se, neste trabalho, adotar o processo de abstração em diferentes níveis. No primeiro nível a realidade é representada num modelo conceitual, que segundo Câmara (1996), nesse estágio, o usuário lida com conceitos mais próximos de sua realidade, minimizando a complexidade envolvida nos diferentes tipos de representação geométrica tais como geo-campos, geo-objetos, objetos cadastrais e objetos não-espaciais.

Para a elaboração do modelo conceitual foi necessário considerar algumas informações arqueológicas a respeito da escolha de locais propícios à ocupação, baseadas em dados obtidos em campo nas etapas de prospecção e escavação, por Thomaz (2002):

“ *... nas confluências dos rios grandes com os rios pequenos, é notória a grande concentração de sítios arqueológicos, sendo explicados pelo fato que os rios de grande porte serviam geralmente para o transporte e pesca e os rios de pequeno porte serviam para banhos e obtenção de água* ...”

“ *... as depressões permitiam a formação de lagoas em épocas de estiagem, confinando grande quantidade de peixes nas suas formações, permitindo a pesca mais acessível, como ainda é notado em algumas regiões do Brasil ...*”

As variáveis importantes no modelo de predição são as observações das presenças ou ausências dos vestígios arqueológicos de uma determinada área representativa no contexto geral, bem como os fatores ambientais associados à ocupação pretérita, ou seja:

- Mapa geológico: representa a formação e a transformação das estruturas envolventes;
- Mapa da geomorfologia: representa as formas do terreno;
- Mapa pedológico: representa as estruturas do solo;
- Mapa altimétrico: representa a altimetria do terreno.

A variável dependente (ausência/presença de sítios arqueológicos) foi verificada em campo durante a etapa de prospecção do Projeto de Salvamento Arqueológico, sendo coletadas as coordenadas geográficas (latitude λ e longitude φ).

Como o modelo de regressão logística proporciona um produto probabilístico, foram inseridas, como novas variáveis, as áreas de influência dos rios de grande e pequeno porte, pois poderiam ser variáveis relevantes na localização das prováveis áreas de ocupações pretéritas. Essas variáveis foram modeladas sobre a base de dados cartográfica, através da ferramenta *buffer*, gerando os seguintes planos de informação:

- *Buffer* das confluências: áreas de confluência dos rios de pequeno porte (Xavantes, Santo Anastácio, ...) com o rio de grande porte (rio Paraná),
- *Buffer* das lagoas: áreas ao entorno das lagoas, vistas principalmente no estado do Mato Grosso (margem direita no sentido montante à jusante);
- *Buffer* de rios grandes: áreas de influência do rio Paraná e
- *buffer* dos rios de pequeno porte: áreas de influência dos pequenos afluentes e rios de pequeno porte.

Todos os intervalos foram estipulados de 1000 em 1000 metros da feição a ser modelada, gerando assim, informações adicionais ao modelo a ser ajustado.

O modelo estático mostrado na figura 16 é basicamente formado por geocampos, pois, como visto acima, os dados que representam a área de pesquisa são formados substancialmente pelo princípio do *planar-enforcement* (restrição de preenchimento do plano): cada localização de um geo-campo está associada a um e somente um valor da variável representada (Câmara, 1996).

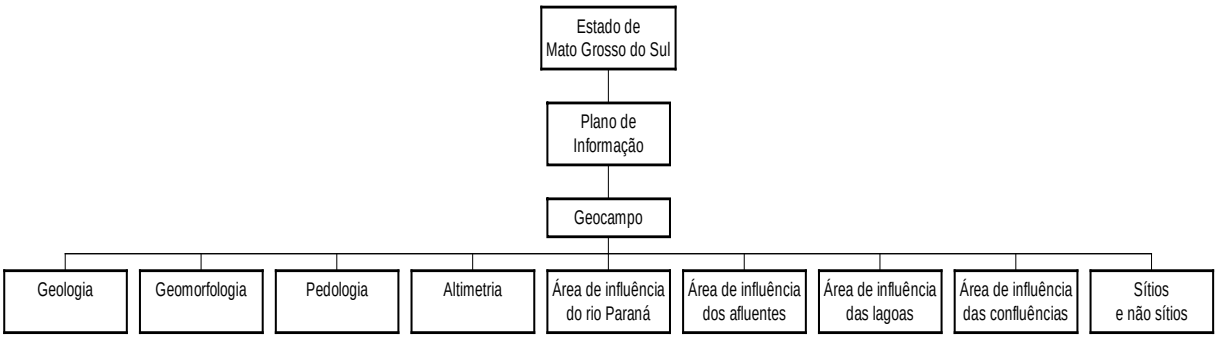


Figura 16 – Diagrama de Temas

5.2.3. Universo de representação

Como o objetivo desse SIG era descrever as prováveis áreas de ocupação pretérita, baseada no modelo de regressão logística, foi necessária a representação dos planos de informação, pois o resultado da aplicação seria um modelo numérico do terreno preditivo, cuja representação é matricial, ou seja, uma função f(x,y) onde a imagem está entre 0 e 1 $\{1 \leq z \leq 0\}$. O modelo a ser gerado deve ser da forma:

Olhando para o modelo acima, pode-se notar que, com a obtenção dos valores β ou de máxima verossimilhança, parâmetros que dão forma à função, cada valor da variável que se encontra em determinada posição, forneceu um valor probabilístico *p(B)* que representa a chance de se encontrar vestígios arqueológicos naquela posição do *grid* (célula).

Por isso, cada plano de informação (PI) representa uma variável X_i e cada posição no *grid* um valor, linha ou coluna associado. O modelo obtido pela regressão logística foi aplicado para cada posição: linha e coluna com os valores das variáveis independentes, armazenados no PI correspondentes estimando um valor probabilístico para cada posição.

Como o modelo trata os dados de forma matricial, com *i* valores dentro de *X* planos de informação, verifica-se a inerente operação com *grids*.

Todos os planos de informação são representados na forma matricial num mesmo domínio, ou seja, num mesmo limite geográfico, onde cada célula (x,y) representa um mesmo tamanho no terreno e cada posição geográfica tem um valor no espaço imagem que representa o valor da variável.

5.3. Banco de Dados Geográficos (DBGs)

Para a geração do Banco de Dados Geográficos no programa Arc/Info, utilizou-se mapas plani-altimétricos (1/10.000); cartas geológicas, pedológicas e geomorfológicas (1/200.000). Foram produzidos os seguintes planos de informação (PIs primários): curvas/cotas; geomorfologia; pedologia; geologia; afluentes; lagoas e rios de grande porte. Todas as feições passaram por uma validação, bem como corrigidos os erros de fechamento de polígonos, topologia e geração dos atributos, para posteriormente serem convertidos em matrizes.

As *coverages* citadas acima permitiram a criação dos PIs secundários, sendo resultantes de processos como o *buffer*, triangulação e interpolação.

A Figura 17, adaptada de Warren (1990), representa bem o procedimento das criações das *coverages* secundárias.

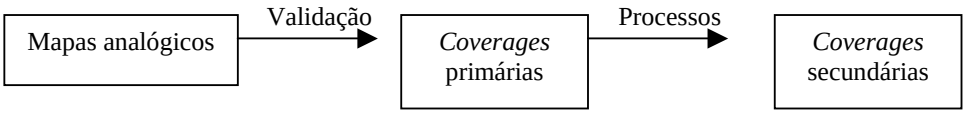


Figura 17: Fluxo generalizado da obtenção das variáveis

As tabelas internas dos planos de informação acima são representadas no dicionário de dados (Capítulo 06). Nesse são descritos as representações das *coverages*, os nomes e tipos dos campos, os valores assumidos e o *datum* geodésico.

Com os planos de informações validados, foi necessária a transformação das mesmas em forma matricial (*grid*), para posterior aplicação do modelo ajustado.

5.4. Ajustando o modelo de Regressão Logística

Para o ajuste do modelo de regressão logística, foi necessário gerar um plano de informação contendo as localizações dos sítios arqueológicos, verificados na etapa de prospecção do Projeto de Salvamento Arqueológico de Porto Primavera, bem como áreas onde não há a presença de vestígios arqueológicos.

O tamanho da amostra (resolução) foi determinado em 15 metros, devido às restrições de *hardware* ou capacidade do disco rígido da estação *RISC* e sendo a menor dimensão de um sítio arqueológico, ou seja, a maioria dos sítios arqueológica da área em questão tem dimensões aproximadas de 10.000 metros quadrados.

Com todos os domínios definidos igualmente aos planos de informação, foram obtidos os valores z das variáveis, nos locais onde existem sítios ou nos locais onde não existem sítios, gerando-se uma tabela de valores apresentada nos resultados.

A partir da planilha gerada, o modelo ajustado foi determinado pela equação do tipo logito, apresentada na fórmula 29.

A método de regressão logística determina os parâmetros β's associados a cada variável, onde a variável só entra no modelo, se for considerada importante no teste de hipóteses, bem como, se passar na tolerância estipulada pelo analista.

No final, tem-se um modelo ajustado, onde somente algumas variáveis são selecionadas para compor o modelo.

5.5. Utilizando o modelo de Regressão Logística

Após o ajustamento do modelo logito à presença de sítios, aplicou-se a equação nas variáveis mais importantes, que foram selecionadas pelo modelo de regressão.

Como todas as *coverages* eram do tipo geo-campo matricial, o modelo pôde ser aplicado nos PI's mais importantes através da álgebra de mapas (ArcGRID), ou seja, para cada célula (*pixel*) tem-se os valores associados aos parâmetros representativos no modelo (β), como sendo pesos para as variáveis.

Posteriormente, os valores das variáveis na porção dessa célula são substituídos no modelo logístico obtido gerando, assim , uma superfície probabilística, onde a imagem da função é dada por valores entre 0 e 1 {0 ≤ valor probabilístico ≤ 1}, como mostra a figura 18.

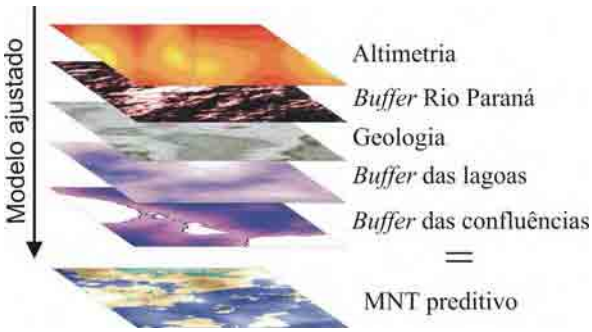


Figura 18: *Grids* das variáveis independentes

Esse processo foi realizado com o recurso de álgebra de mapas do Arc/Info, que produziu um novo PI representando a probabilidade de encontrar vestígios arqueológicos em cada uma das células do MNT (*lattice*).

O esquema a seguir, mostra o processamento das *coverages*, até o respectivo resultado.

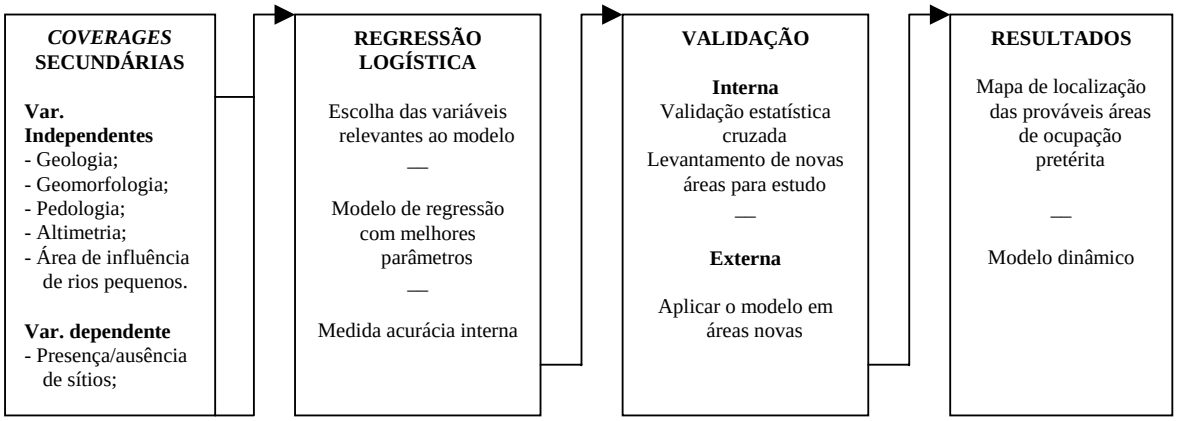


Figura 19: Esquema da modelagem preditiva (Adaptado Warren, 1990)

No próximo capítulo apresenta detalhadamente o processo de ajustamento do modelo logístico, onde as variáveis utilizadas são selecionadas e submetidas a uma validação de hipóteses.

O mapa obtido pela aplicação do modelo ajustado é apresentado, bem como uma discussão dos resultados obtidos.

6. RESULTADO

A aplicação e o estudo do método de regressão logística, para dados provenientes do Projeto de Salvamento Arqueológico de Porto Primavera, resultou um modelo numérico preditivo. Esse mapa pode ser uma ferramenta na qual o arqueólogo pode utilizar como indicador para a procura de materiais em áreas ainda não levantadas, otimizando uma das etapas de um Projeto de Salvamento Arqueológico: a prospecção.

Porém, deve ser ressaltado, como a proposta de Bona (1992), que o modelo pode ser aprimorado, no sentido de novas observações ou variáveis serem inseridas para uma melhor aproximação da preferência das antigas civilizações, de acordo com as análises dos resíduos da determinação do modelo preditivo.

Nesse capítulo será apresentado o resultado da modelagem do banco de dados geográficos, bem como a análise da regressão logística e o modelo numérico preditivo do terreno.

6.1. Banco de Dados Geográficos

O processamento para determinação do modelo de regressão logística para a área de estudo teve que ser realizado sob a forma matricial ou *grids*, partindo de um mesmo domínio das funções $f_i(x,y)$ com i=1, ..., n e n = número de variáveis.

O domínio para a aplicação de planos de informação matriciais com modelo logístico foi definido, pois o resultado é um novo plano de informação probabilístico, onde esse está presente no mesmo domínio e o valor de cada célula tem valor probabilístico entre 0 e 100%.

De acordo com os capítulos anteriores, na regressão logística utiliza-se variáveis independentes, que explicam a variável dependente, a presença/ausência de ocorrências arqueológicas. Essas variáveis são representadas pelas características físicas do terreno e informações decorrentes do universo arqueológico, ou seja, as características das preferências dos índios para ocupações de determinados locais.

No diagrama temático visto no capitulo anterior, foram apresentadas as variáveis ambientais e as variáveis que foram modeladas com base na consulta aos arqueólogos, sendo representadas em escalas nominais, ordinais, intervalos ou variações em escala.

As variáveis ambientais que constituem a base cartográfica são:

- Mapa geológico: representa a origem, a composição e a transformação das estruturas da Terra em escala nominal (Figura 20);
- Mapa de geomorfologia: representa as formas e a evolução do relevo, sendo representada na escala nominal (Figura 21);
- Mapa pedológico: representa os tipos de solo, cuja escala é nominal (Figura 22) e
- Mapa altimétrico: representa a altimetria do terreno, em escala ordinal (Figura 23).

As variáveis implementadas com base no conhecimento dos arqueólogos foram geradas a partir da ferramenta *buffer*: áreas de influência de uma determinada feição. Essas áreas foram consideradas em intervalos de 1000 em 1000 metros, pois a escolha da distância decorreu do tamanho considerado de um sitio arqueológico, que pode atingir 1000 metros de extensão.

As variáveis geradas a partir do Universo Arqueológico são todas ordinais:

- Áreas de influências das confluências: áreas com base na distância das confluências dos rios de pequeno porte com o rio Paraná (Figura 24);
- Áreas de influências das lagoas: áreas com base na distância das lagoas (Figura 25);
- Áreas de influências do Rio Paraná: áreas com base na distância do leito do rio Paraná (Figura 26) e
- Áreas de influências dos rios de pequeno porte: áreas com base na distância dos leitos dos rios de pequeno porte (Figura 27).

O dicionário do banco de dados geográficos demonstra as categorias existentes no projeto, suas respectivas classes e representações (Tabela 03).

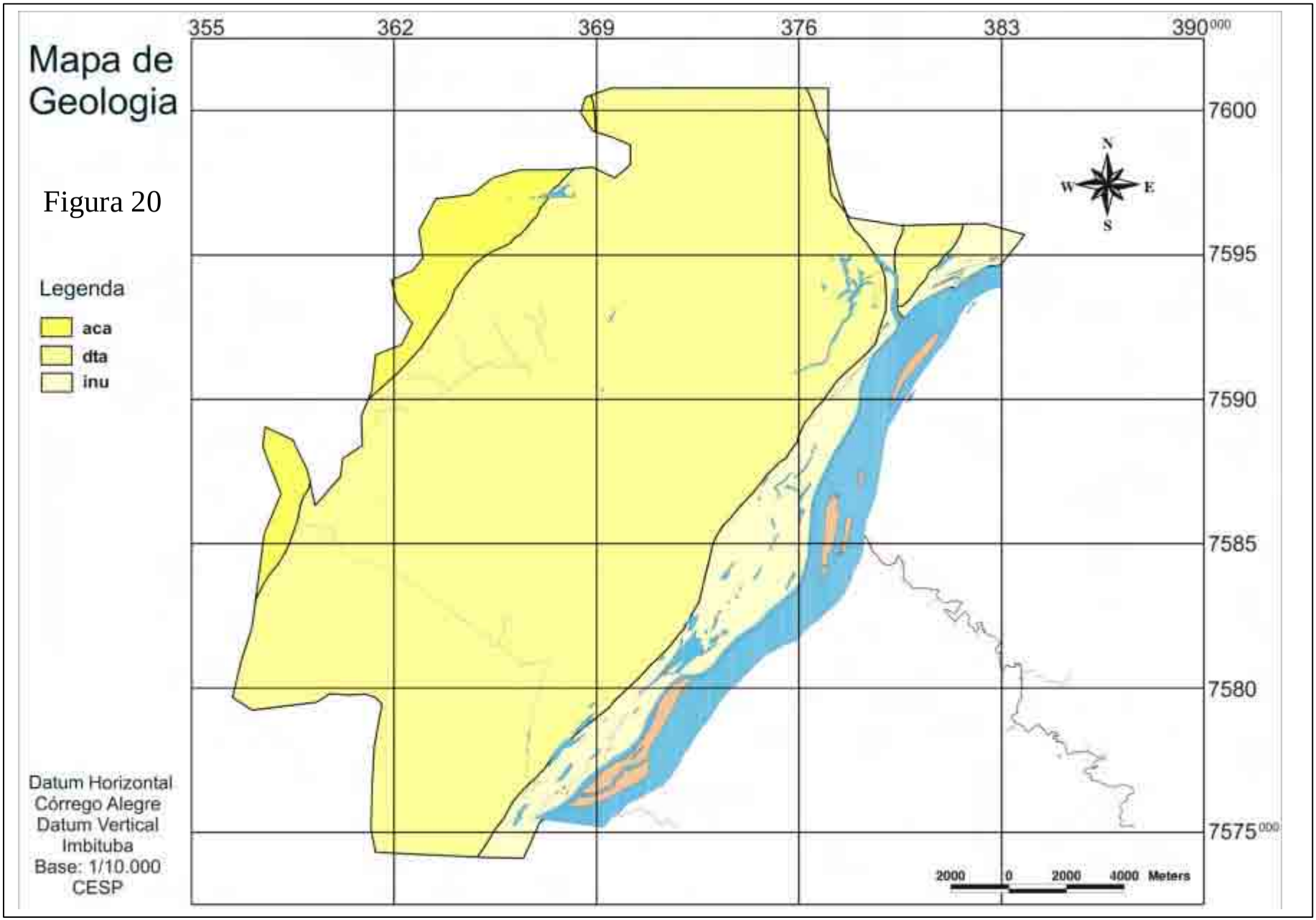
Tabela 03: Dicionário de dados

Categoria	Descrição	Modelo	Campos	Tipo	Classes	Cores
Geologia	Representa a gêneses do terreno	Temático	Code - Cobertu	Binário *Char* (10,10)	Pi = 10 Dta = 01 Aca = 11 Aba = 00	aca dta inu
Geomorfologia	Representa a conformação do terreno	Temático	Code – Cobertu	Binário *Char* (10,10)	Ta1 = 100 Pa = 111 Ta2 = 011 Sac = 001	pal sac ta1 ta2
Pedologia	Representa a composição do terreno	Temático	Code – Cobertu	Binário *Char* (10,10)	1 – AL2 2 - AL2 + GPH 3 - AQD + AL2 + GPH 4 - GH 5 - GH + GPH 6 - GPH 7 - LV3 8 - LVIA 9 - PV1	al+gph al2 al2+gph aqd+al2+gph gh gh+gph gph
Altimetria	Representa a altimetria da área	Temático	Code – Dist.	Binário *Float* (4,12,3)	1 – 255 –258 2 – 258 – 261 3 – 261 – 264 4 – 264 – 267 5 – 267 – 270 m	255 270 m
Área de influência das confluências	Representa área de influência	Temático	Code – Dist	Binário Inteiro	1000 2000 3000 4000 5000 6000 m	1000 m 2000 m 3000 m 4000 m 5000 m 6000 + m
Área de influência das margens do rio Paraná	Representa área de influência	Temático	Code – Dist	Binário Inteiro	1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000 16000 m	1000 m 2000 m 3000 m 4000 m 5000 m 6000 m 7000 m 8000 m 9000 m 10000 m 11000 m 12000 m 13000 m 14000 m 15000 m 16000 m

Área de influência dos afluentes	Representa área de influência	Temático	Code – Dist	Binário Inteiro	1000 2000 3000 4000 5000 + 6000 metros	1000 m 2000 m 3000 m 4000 m 5000 m 6000 m
Localização dos sítios e não sítios	Representa área de influência	Temático	Code – Dist	Binário Inteiro	1 1	
Área de influência das lagoas	Representa área de influência	Temático	Code – Dist	Binário Inteiro	1000 2000 3000 4000 5000 6000 m	1000 m 2000 m 3000 m 4000 m 5000 m + 6000 m

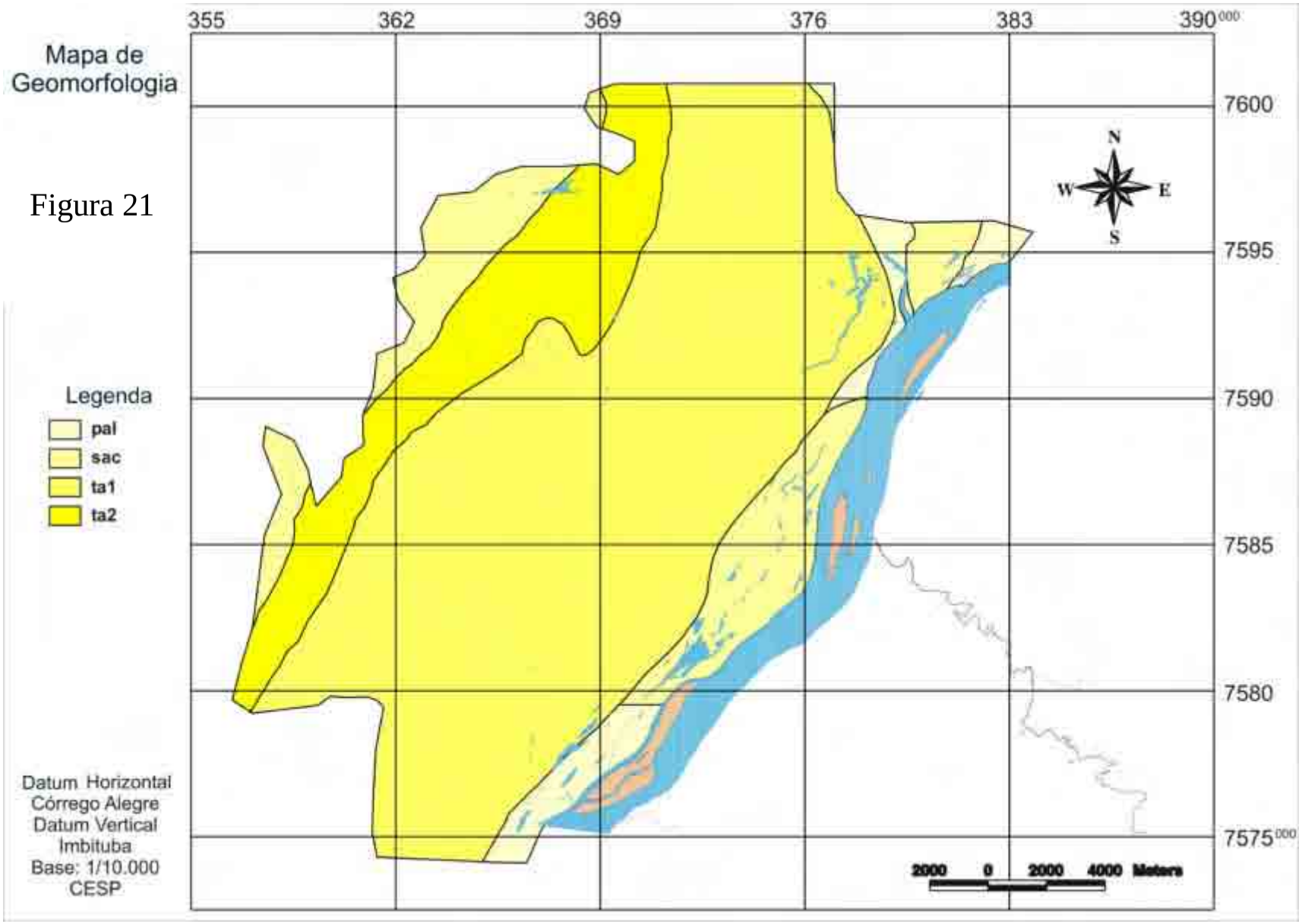
Tabela 04: Nomenclatura das classes

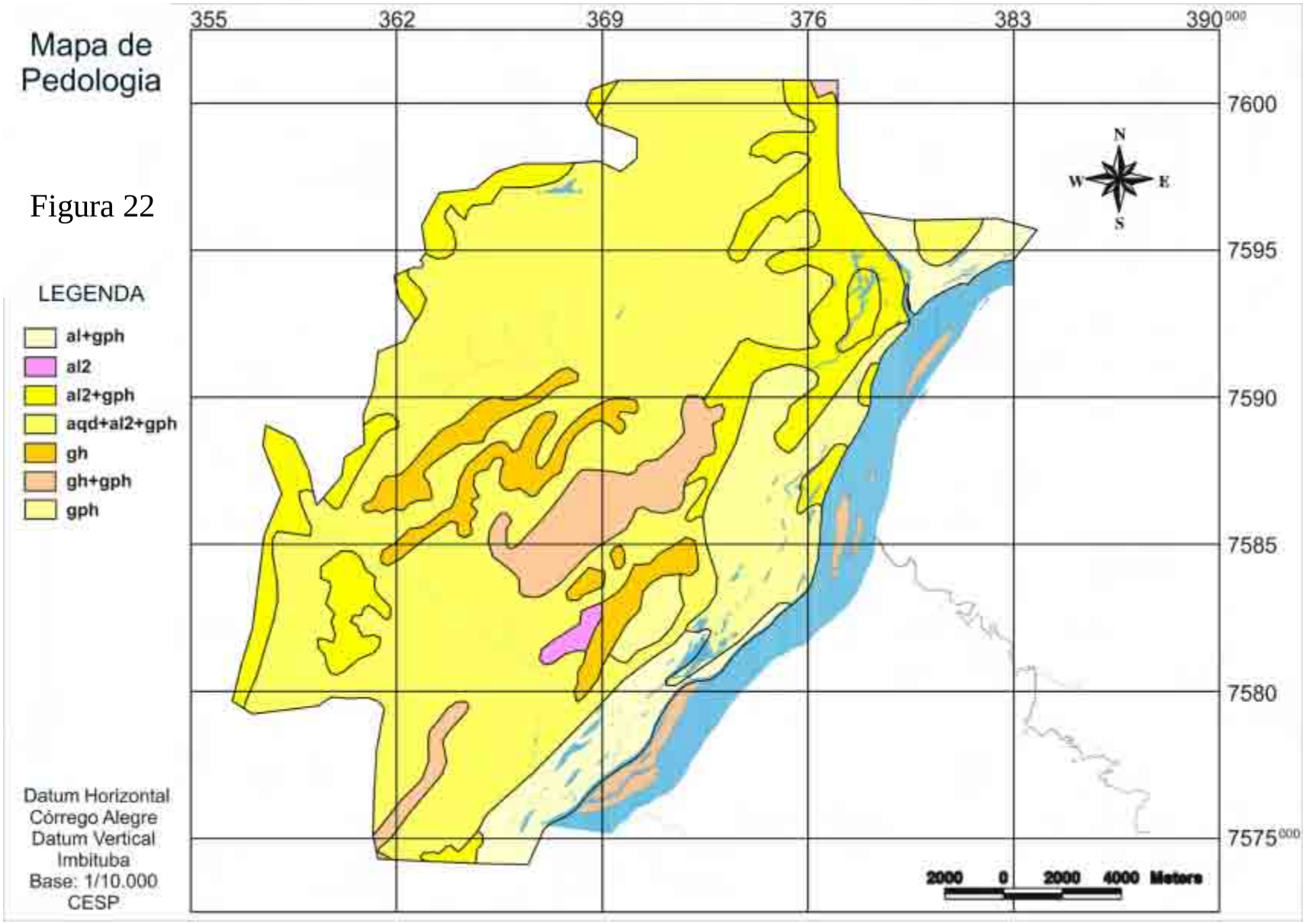
Categoria	Nomenclatura das classes
Geologia	Arenito Bauru, arenito Caiuá, planície de inundação e depósito de terraço aluvionar
Geomorfologia	Planície aluvionar, terraço aluvionar I, terraço aluvionar II e superfície aplainada de Caiuá, Bauru e basalto com drenagem conseqüente.
Pedologia	Latossolo vermelho escuro textura média a arenosa acentuadamente drenado, relevo suave ondulado (LVia) Latossolo vermelho escuro bem drenado textura argilosa, relevo suave ondulado (LV3) Podzólico vermelho amarelo bem drenado, textura média , relevo praticamente plano (PV1) Solos aluvionares, textura média imperfeitamente drenado relativa. Praticamente plano + gley pouco úmido, textura argilosa imperfeitamente mal drenado, relevo plano de várzea. (AL2+GPH) Solos gley hidromórficos textura argilosa, muito mal drenado, relevo plano de várzea. (GH) Solos gley hidromórficos. Textura argilosa muito mal drenado, relevo plano de várzea mais gley pouco húmico, textura argilosa imperfeitamente a mal drenado relevo plano de várzea (GH+GPH). Gley pouco húmico, textura argilosa imperfeitamente a mal drenado relevo plano de várzea (GPH). Solos aluviais textura média imperfeitamente drenado, relevo praticamente plano. (AL2). Areia quartzoza vermelho amarelo, distrof. moderadamente drenado, relevo praticamente plano + solo aluvionar, textura média imperfeitamente drenado, relevo praticamente plano + gley pouco húmico, textura argilosa imperfeitamente mal drenado relevo plano de várzea. (AQd + AL2 + GPH).


Mapa de Geologia
Figura 20
Legenda
aca
dta
inu
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
2000
0
2000
4000
Meters

Mapa de Geomorfologia

Figura 21




Mapa de Pedologia
Figura 22
LEGENDA
al+gph
al2
al2+gph
aqd+al2+gph
gh
gh+gph
gph
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
2000
0
2000
4000
Meters

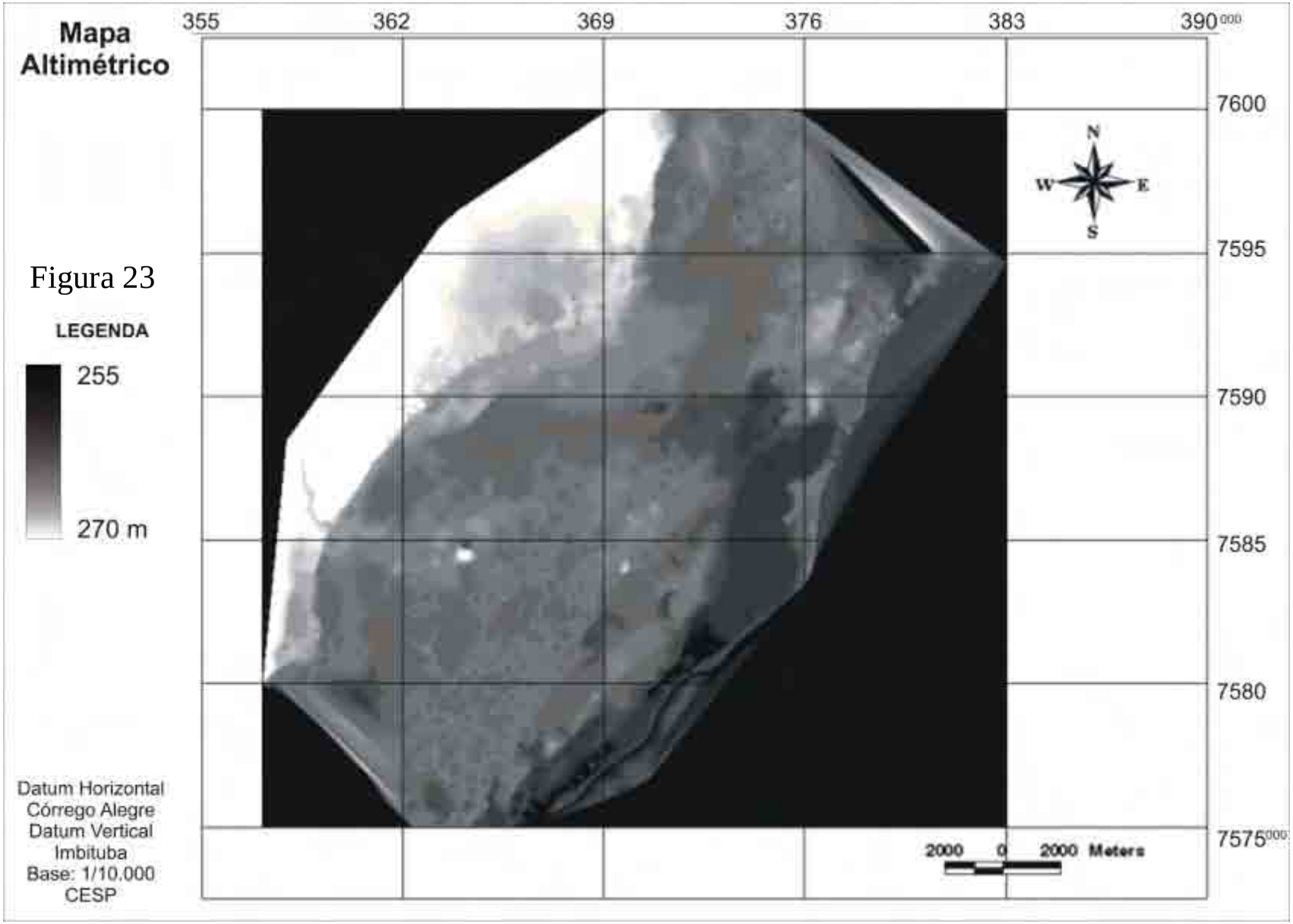
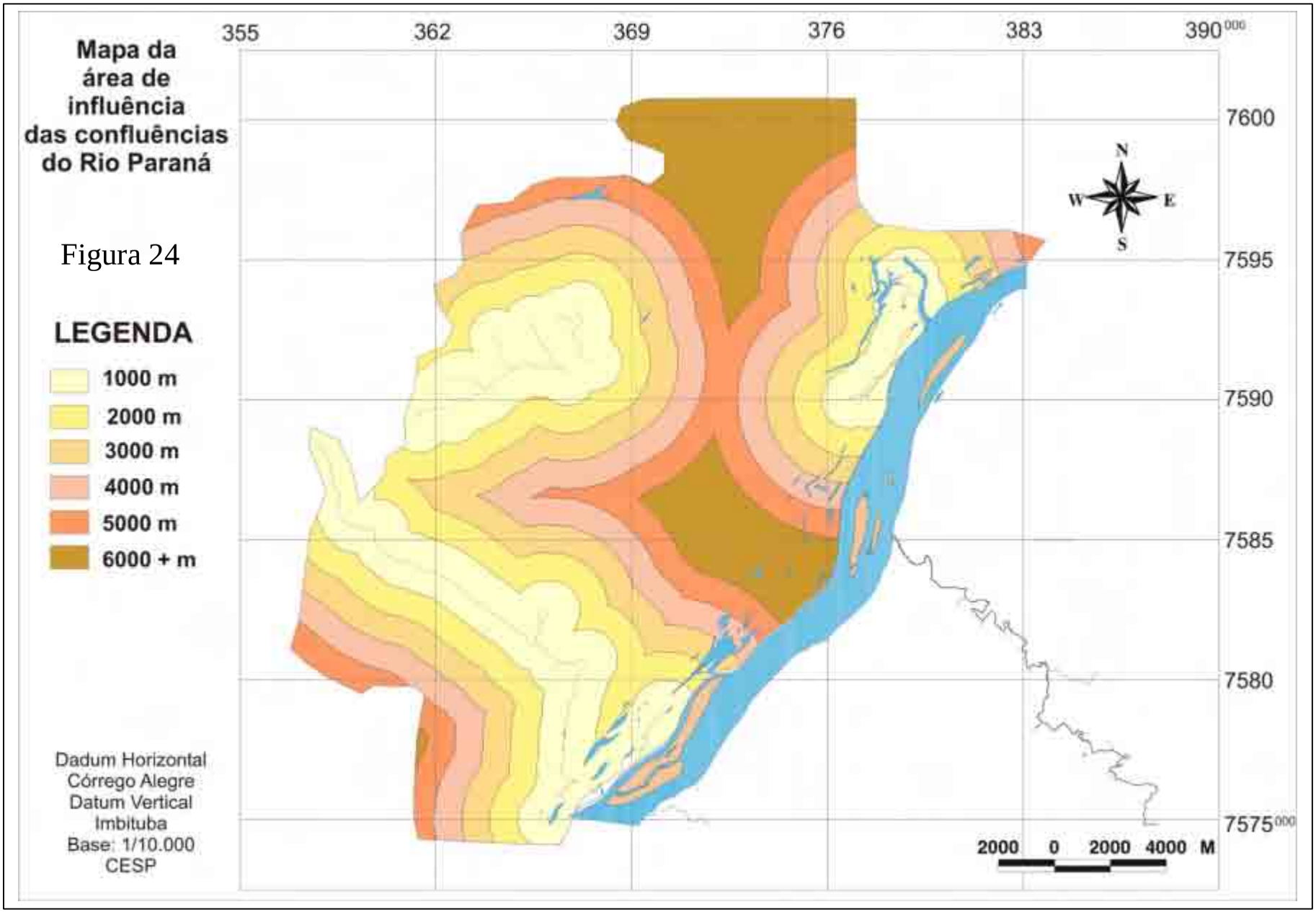
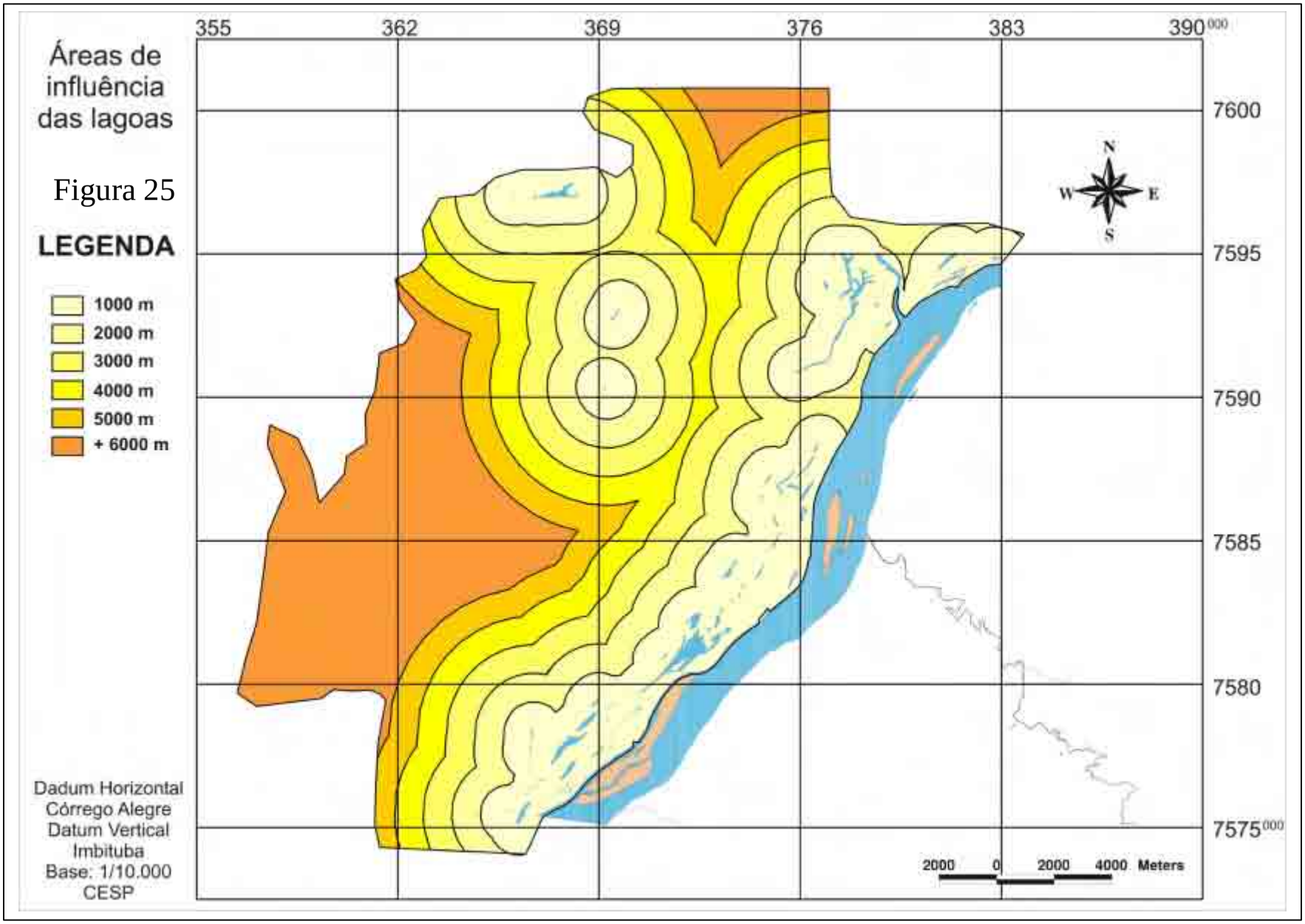
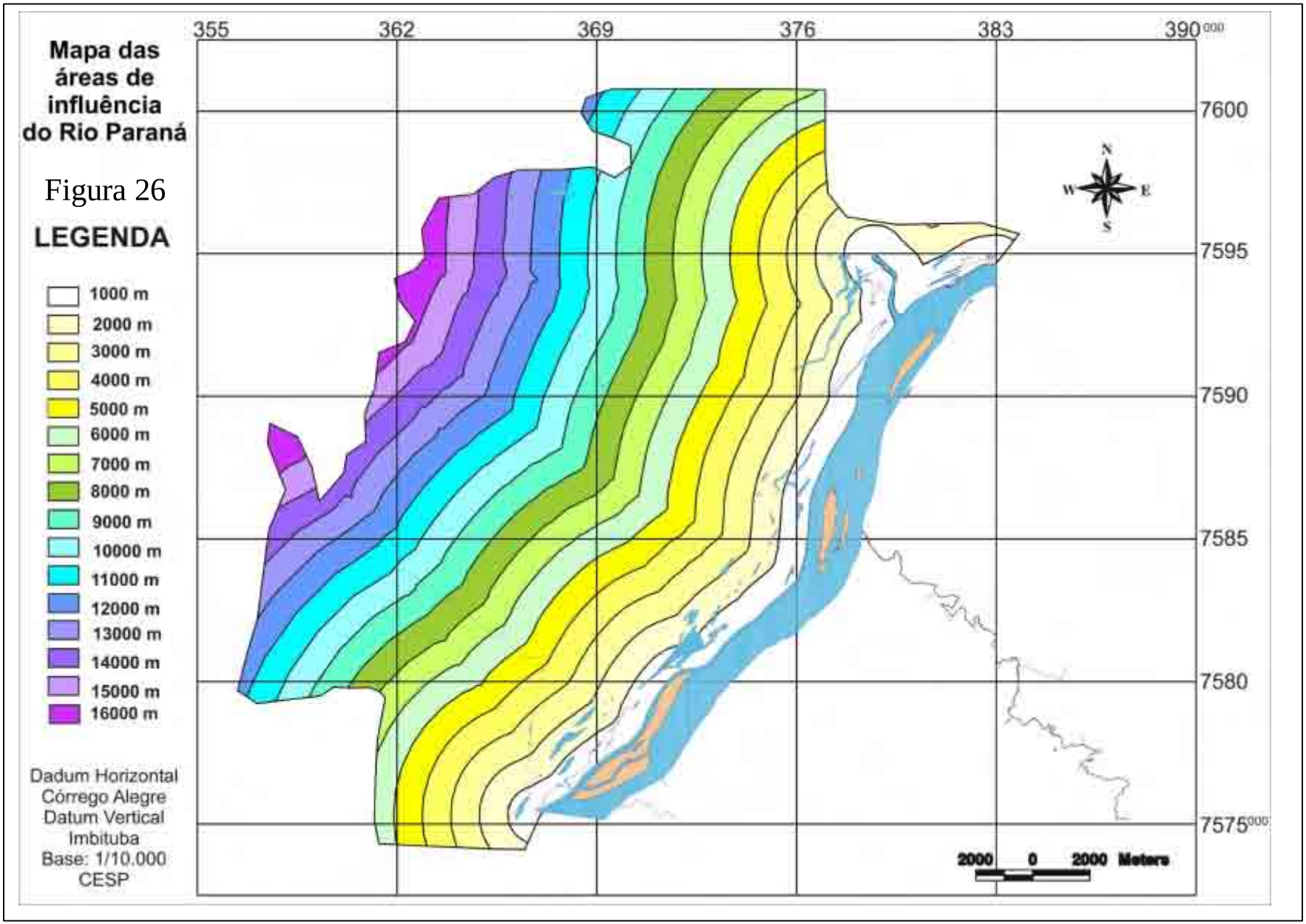
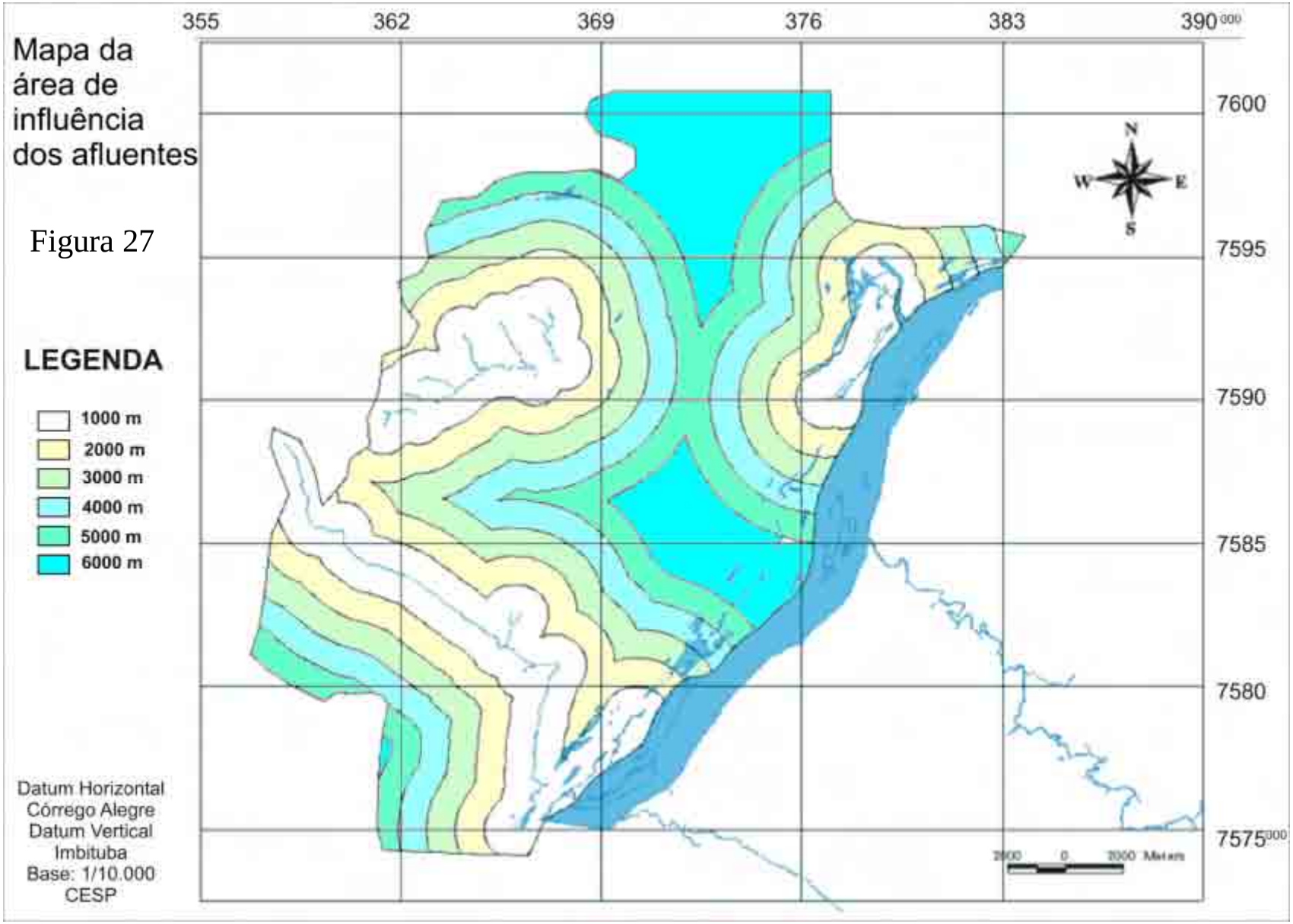

Mapa Altimétrico
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
LEGENDA
255
270 m
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
2000 0 2000 Meters

Figura 23


Mapa da área de influência das confluências do Rio Paraná
Figura 24
LEGENDA
1000 m
2000 m
3000 m
4000 m
5000 m
6000 + m
Dadum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
2000 0 2000 4000 M


Áreas de influência das lagoas
Figura 25
LEGENDA
1000 m
2000 m
3000 m
4000 m
5000 m
+ 6000 m
Dadum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
2000
0
2000
4000
Meters


Mapa das áreas de influência do Rio Paraná
Figura 26
LEGENDA
1000 m
2000 m
3000 m
4000 m
5000 m
6000 m
7000 m
8000 m
9000 m
10000 m
11000 m
12000 m
13000 m
14000 m
15000 m
16000 m
Dadum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
2000 0 2000 Meters


Mapa da área de influência dos afluentes
Figura 27
LEGENDA
1000 m
2000 m
3000 m
4000 m
5000 m
6000 m
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
2000
0
2000 Meters

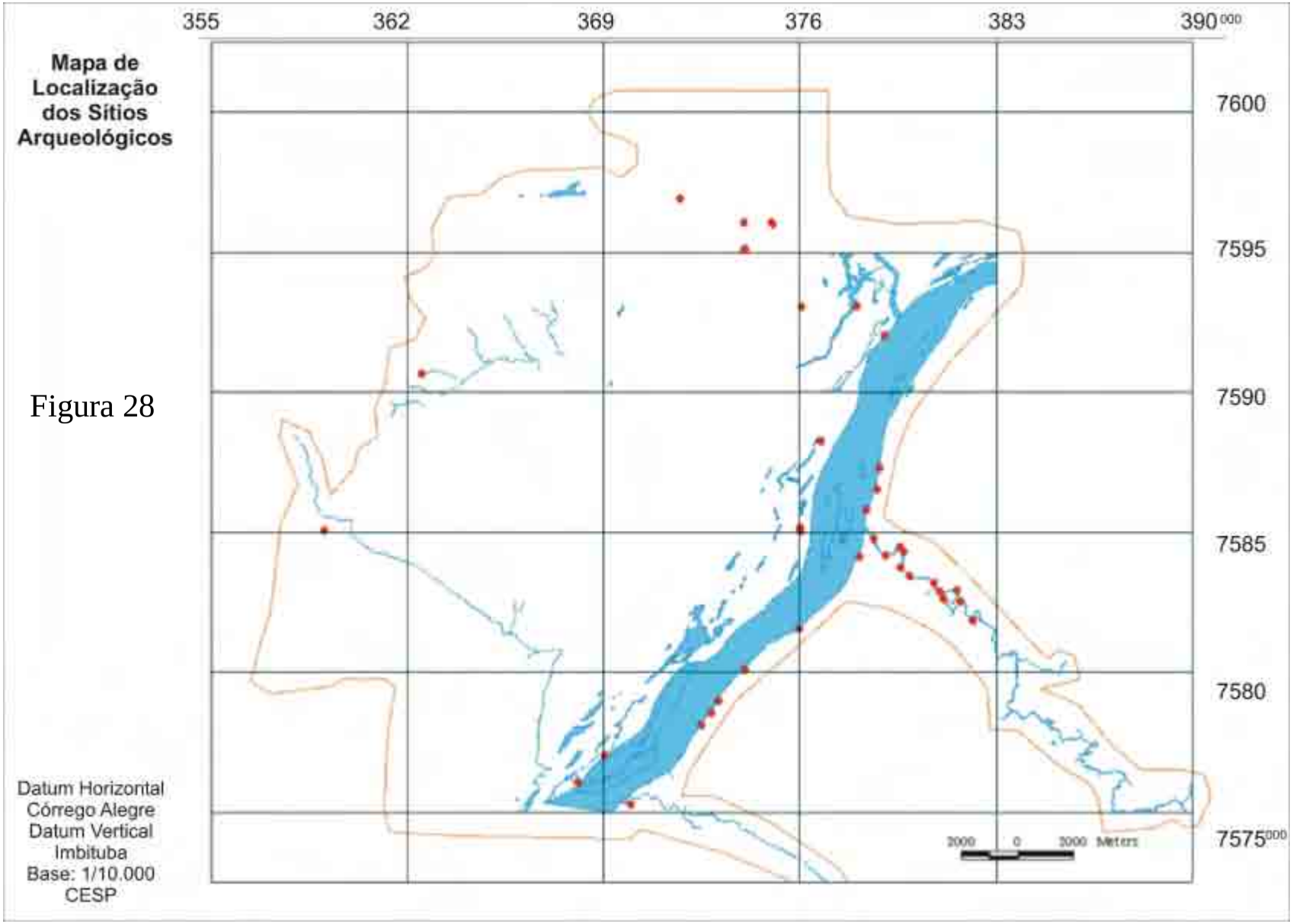

Mapa de Localização dos Sítios Arqueológicos
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
2000
0
2000 Meters
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP

Figura 28

A variável dependente (ausência/presença de sítios arqueológicos) foi verificada em campo, durante a etapa de prospecção do Projeto de Salvamento Arqueológico de Porto Primavera, sendo realizada por pesquisadores da Universidade Federal de Mato Grosso.

De todas as variáveis presentes no modelo estático, somente algumas foram aceitas como mais importantes no processamento do modelo de regressão logística, fato esse, denotado pela submissão ou equiparação de variáveis semelhantes ou irrelevantes no processo de ocupação das populações que ocuparam a região. Como exemplo pode-se citar que a variável geomorfologia pode ser substituída pela altimetria, onde ambas apresentam a conformação do terreno, tendo a mesma importância para a regressão.

A escolha das variáveis mais importantes foi realizada por passos. Como se trata de um processo iterativo de determinação de parâmetros, a cada passo insere-se uma variável no modelo onde são realizados os testes dos valores observados, relacionados com os estimados.

6.2. Ajustando o modelo de regressão logística

Antes de serem processados os dados do projeto, os limites para adicionar ou remover variáveis foram configurados. Desta forma para retirada da variável o valor *p* ou *p*-valor deve ser maior que 0.1500 e para entrada da variável o valor *p* deve ser menor que 0.1000.

No primeiro passo foi determinado o logaritmo da verossimilhança para o modelo contendo o termo constante (β_0), considerando que nenhuma variável explicativa fornece informações significativas no modelo.

Usando um pacote estatístico de regressão logística, obteve-se os seguintes resultados apresentados na Tabela 05.

Tabela 05- Resultados do ajustamento do modelo contendo apenas o termo constante.

Termo	coeficiente	Erro Padrão (ep)	coef/e.p.
Constante	-1.3471	0.1392	-9.677
Log da verossimilhança = -160,360			

As estimativas de β_0 é $\hat{\beta}_0$ = -1,3471 e seu erro padrão é igual a 0,1392. O teste de Wald pode ser realizado comparando-se $\hat{\beta}_0$ com a estimativa do seu erro padrão. O resultado

da razão entre esses dois valores, sob a hipótese que $\beta_0 = 0$, terá uma distribuição normal padrão. Com um valor extremamente pequeno (-9,677) não há evidência para a hipótese ser aceita.

No segundo passo, foram ajustados os possíveis modelos univariados de regressão logística (um para cada variável explicativa), cujos resultados estão apresentados na Tabela 05 e comparados seus respectivos logaritmos de verossimilhança. Seleciona-se a variável mais "importante" com menor *p-valor*. A Tabela 06 mostra que duas variáveis apresentaram p-valores iguais a 0,0000 (não é igual a zero, existem dígitos significativos abaixo de 10^{-4}) e log da verossimilhança próximos, mas a altimetria tem apenas um grau de liberdade. Então a variável mais importante é altimetria. A importância dessa variável no modelo foi verificada através do teste da razão de verossimilhança e o teste de Wald.

O terceiro passo inicia-se com um ajuste do modelo de regressão logística contendo a variável altimetria. Para verificar se outras variáveis explicativas são importantes, uma vez que altimetria está no modelo, ajustam-se modelos de regressão logística contendo altimetria e cada uma das variáveis explicativas. Nesse ponto é visualizada a estatística para adição ou remoção dos termos, sendo verificada pelo *p-valor*, que condiz com a probabilidade da variável assumir um valor maior ou igual ao valor observado. A seleção dessa variável é feita como descrito anteriormente no segundo passo. O valor para aceitação da variável (entrada) tem que ser menor que o valor atribuído como limite probabilístico. Para verificar se uma variável que tenha sido acrescentada ao modelo não é mais importante, é necessário comparar seu *p-valor* com o limite probabilístico fixado previamente. A variável é eliminada do modelo se seu *p-valor* for maior que esse limite.

Tabela 06 – Logaritmo da verossimilhança, graus de liberdade (g.l.) e p-valores para as variáveis entrarem ou permanecerem no modelo.

Variável	χ^2 aprox entrada	g.l .	χ^2 aprox remoção	g.l.	*p-valor*	log veros.
buffRP	10.53	1			0.0012	-155.0965
buffaflu	0.98	1			0.3214	-159.8684
buffconf	3.77	1			0.0521	-158.4730
lagoas	1.80	1			0.1794	-159.4588
geologia	24.96	1			0.0000	-147.8815
geomorf	0.11	1			0.7433	-160.3064
pedolog	30.21	4			0.0000	-145.2526
altimetria	32.21	1			0.0000	-144.2530
constante			115.96	1	0.0000	-218.3412

Logaritmo da verossimilhança = -144.253

Os passos subseqüentes seguem o mesmo padrão de análise, realizando iterações até a última variável.

A Tabela 07 mostra as variáveis mais importantes para o modelo. Os valores em **negrito**, mostram que os valores são menores que o limite probabilístico, ou seja, são mais significativas na explicação do modelo de presença de sítios arqueológicos. As variáveis mais importantes para o modelo foram *Buffer* do Rio Paraná, *Buffer* das confluências, *Buffer* das lagoas e Altimetria.

Tabela 07: Logaritmo da verossimilhança, graus de liberdade (g.l.) e p-valores para as variáveis entrarem ou permanecerem no modelo.

Variável	χ^2 aprox entrada	g.l .	χ^2 aprox remoção	g.l.	*p-valor*	log veros.
BuffRP			62.39	1	**0.0000**	-77.8820
Buffaflu	0.20	1			0.6551	-46.5889
BuffCon			4.41	1	**0.0357**	-48.8951
Lagoas			14.68	1	**0.0001**	-54.0288
Geologia			29.51	1	**0.0000**	-61.4446
Geomorf	2.64	1			0.1043	-45.3697
Pedologia	1.27	4			0.8658	-46.0517
Altimetria			182.25	1	**0.0000**	-137.8119

Tabela 08: Resultados do ajuste dos parâmetros

Variáveis	Coeficiente	Padrão erro	Coef/E.p.
Buffrg	-0.19509E-02	0.3330E-03	-5.858
Buffcon	0.50166E-03	0.2454E-03	2.044
Lagoas	0.75441E-03	0.2181E-03	3.459
Geologia	4.5896	1.052	4.364
Altimetria	6.3859	0.9556	6.683
Constante	-14.557	2.158	-6.745

O modelo de regressão logística para predição de ocorrência de vestígios arqueológicos é dado por:

$$p(B) = \frac{e^{-14.557+4.5896geo\log_i + -0.0019509bufferRP_i + 0.00075441bufferlagoas_i + 0.00050166bufferconfluencias_i + 6.3859altimetria_i}}{1+e^{-14.557+4.5896geo\log_i + -0.0019509bufferRP_i + 0.00075441bufferlagoas_i + 0.00050166bufferconfluencias_i + 6.3859altimetria_i}}$$

Com base nas ferramentas de álgebra de mapas do Arc/Info (GRID), o modelo ajustado foi aplicado em todos os planos de informação, onde pode-se realizar as operações básicas (multiplicação, adição, subtração e divisão) em cada posição.

Cada *cover* estava representada no formato matricial (*grid*), onde cada célula representa o mesmo tamanho no mesmo domínio.

Como resultado, foi gerado um modelo numérico do terreno de valores probabilísticos (Figura 29), que corresponde ao mapeamento probabilístico entre intervalo 0 a 100%.

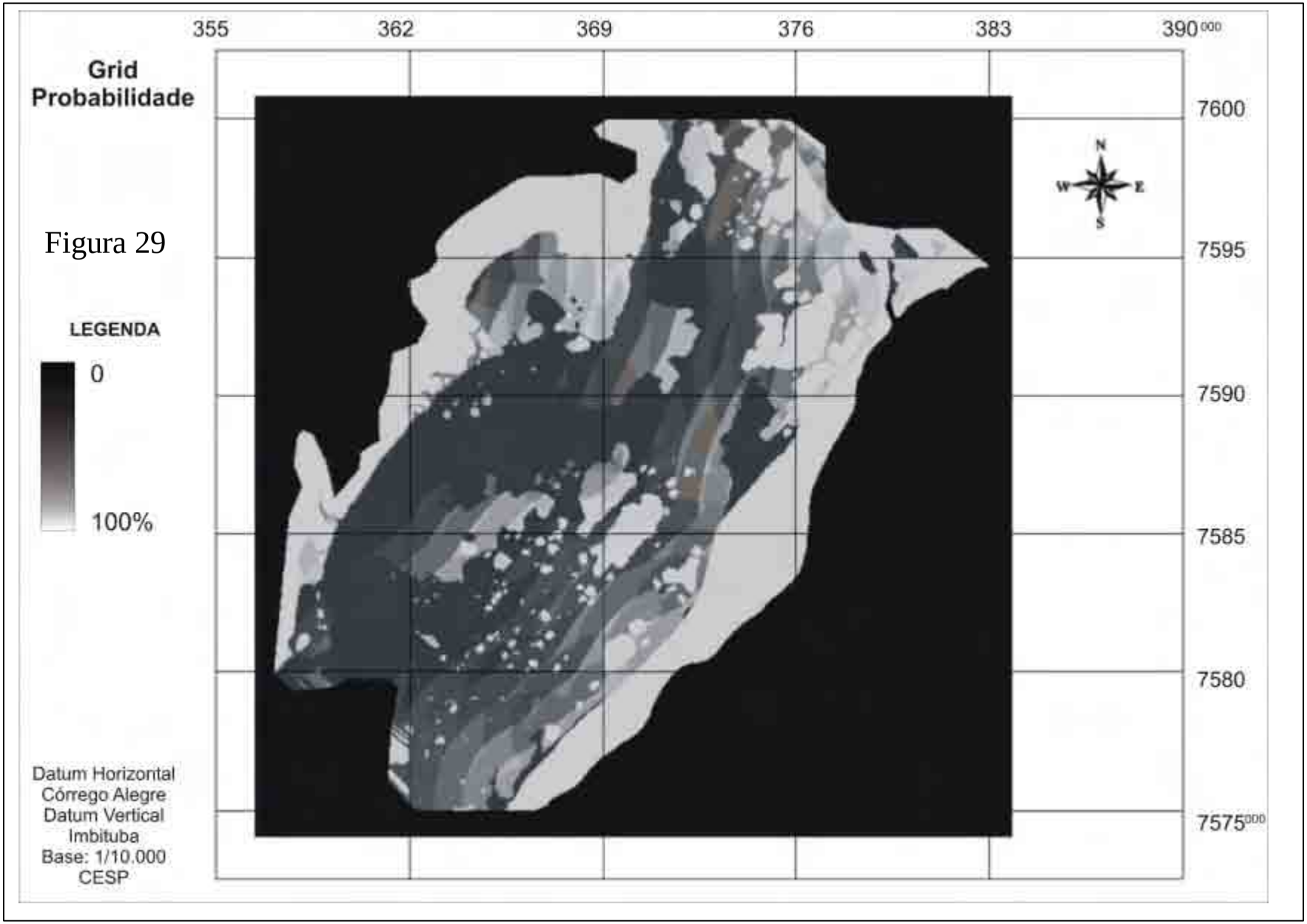
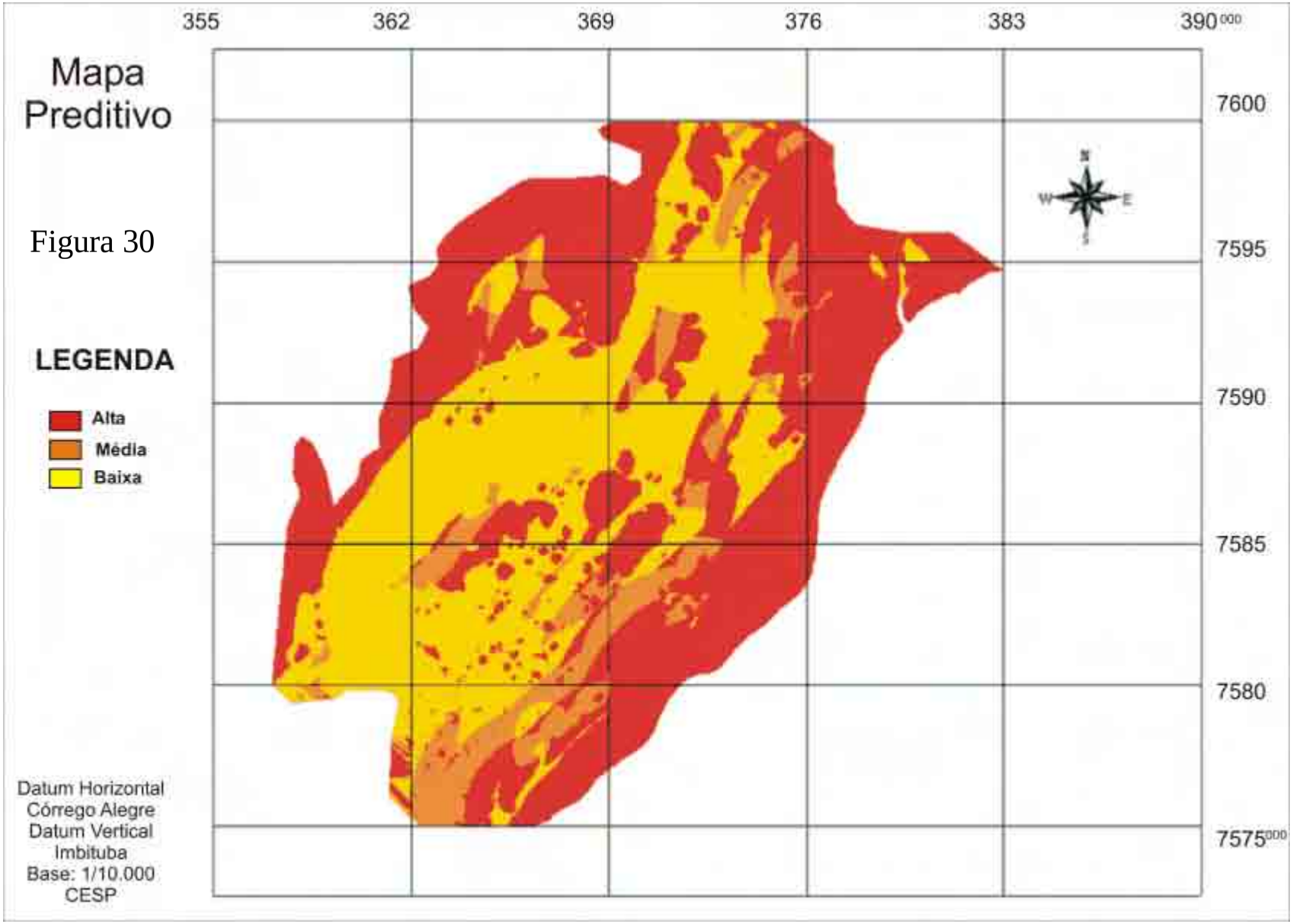

Grid
Probabilidade
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S
LEGENDA
0
100%
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP

Figura 29

A partir da superfície probabilística realizou-se uma classificação entre isovalores, utilizando-se três categorias iguais. As classes foram determinadas como alta probabilidade (66% até 100%); média probabilidade (33% até 66%) e baixa probabilidade (0% até 32%), como mostra a figura 30.


Mapa
Preditivo
Figura 30
LEGENDA
Alta
Média
Baixa
Datum Horizontal
Córrego Alegre
Datum Vertical
Imbituba
Base: 1/10.000
CESP
355
362
369
376
383
390000
7600
7595
7590
7585
7580
7575000
N
W
E
S

Assim, com o mapa classificado, a equipe de prospecção arqueológica, pode realizar a verificação dos locais com probabilidades médias e altas, sendo descartadas outras áreas, que foram consideradas irrelevantes na presença de sítios arqueológicos.Ou seja, com equipamentos de localização (GPS) podem ser verificadas áreas cujas características não fogem do princípio conhecido dos arqueólogos, e podem ser afirmados com determinada precisão, a razão pela qual os locais são vistoriados, não sendo mais necessário um sistema informal para a etapa de prospecção em Projetos de Salvamento Arqueológico.

7. CONCLUSÃO

Como visto nos capítulos anteriores, em Projetos de Salvamento Arqueológico, o arqueólogo realiza a busca de vestígios em todo o terreno a ser abrangido por um determinado empreendimento.

Após a constatação da existência de materiais arqueológicos e a obtenção do seu posicionamento, faz-se a escavação das áreas determinadas, denominadas sítios arqueológicos.

A proposta da aplicação do modelo preditivo de regressão logística, decorre da redução da etapa de verificação em campo, pois esse processo é oneroso, às vezes superficial e um tanto impreciso.

Na aplicação desse método, a etapa de prospecção resume-se a uma amostragem significativa de alguns locais de sítios e não sítios, bem como as variáveis que permitem descrever a característica dessa ocupação, assim sendo, torna-se necessário ter uma base de dados cartográfica da área a ser afetada pelo empreendimento e uma amostragem casual simples.

Verificou-se que o modelo logístico ajustado fornece a probabilidade de sucesso na obtenção de locais arqueológicos e a partir dos parâmetros ajustados, é possível predizer os locais ainda não prospectados, confiando-lhes uma determinada precisão devido aos testes de hipóteses aplicados durante a aplicação do modelo logístico.

Assim sendo, a vistoria da área torna-se necessária somente na etapa da amostragem casual, onde são obtidas as amostras da variável resposta (Figura 31)

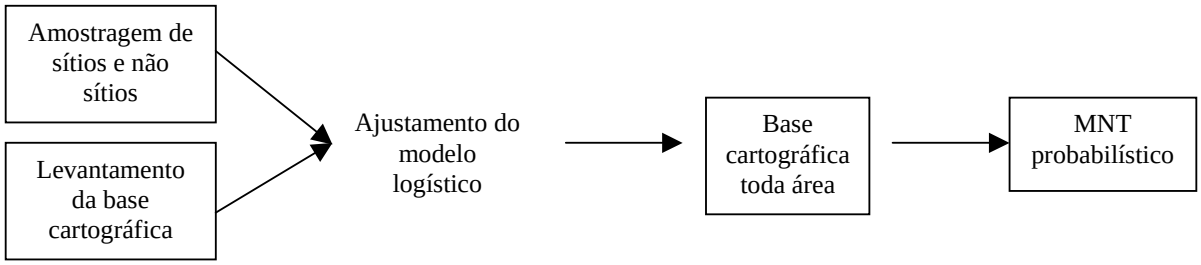


Figura 31: Aplicação do modelo logístico

As ferramentas de Geoprocessamento juntamente com o método de regressão logística permitiram o desenvolvimento do processo de prospecção, viabilizando um método probabilístico que pode predizer espacialmente as prováveis áreas de ocupação pretérita, podendo predizer novos dados não prospectados.

Algumas variáveis foram selecionadas pelo método de regressão logística para explicar o sucesso de encontrar vestígios arqueológicos e forneceram algumas informações conhecidas, permitindo indicar algumas condições da paisagem e costumes das populações pretéritas, a seguir serão descritas as variáveis e a aproximação no universo passado:

- altimetria: essa variável foi escolhida como importante, pois as ocupações davam-se, na maioria das vezes, em locais altos, onde não havia influência em épocas de grande precipitação;
- proximidades do Rio Paraná: esse fator foi relevante, pois a grupo tupiguarani utilizava o rio Paraná para locomoção e pesca;
- proximidades das lagoas: a área da pesquisa é uma planície de inundação, onde eram constantes as inundações, assim sendo, os peixes ficavam retidos nas lagoas e na estiagem ficavam à vista, sendo presas fáceis para a captura, como nota-se, até hoje, em algumas tribos da Amazônia;
- proximidades das confluências: como as pesquisas nessa área já indicaram, os indígenas utilizavam os rios pequenos para o banho e coleta de água, e no rio Paraná era utilizado para pesca e navegação, então nota-se a importância da confluência dos rios. Normalmente na frente das confluências formam-se ilhas, devido à dinâmica dos fluídos, nesse trabalho, foi notado que existem sempre sítios na confluência, na ilha e no lado posterior à ilha, podendo ser admitido a preferência para travessia nesses pontos e principalmente por existirem corredeiras, sendo assim locais de fácil obtenção de pesca e
- geologia: essa variável entrou no modelo pois a formação do terreno explica grande parte da matéria prima dos vestígios encontrados. A figura 32 mostra um lítico lascado encontrado às margens do Rio Paraná, esse material segundo geólogos é formado por um processo de metamorfose de

sedimentos, no qual ocorre agrupamento do arenito, determinado como silicificação. Esse tipo de material não aflora na estrutura geológica da área de estudo, porém pode-se indicar, pelo formato da rocha e o pelo polimento, é que esse material veio sendo arrastado pela correnteza rio abaixo e veio ser depositado em camada sobreposta por material sedimentar. Na medida em que os canais meândricos começavam a erodir as margens, ocorria o afloramento e deposição desse material.

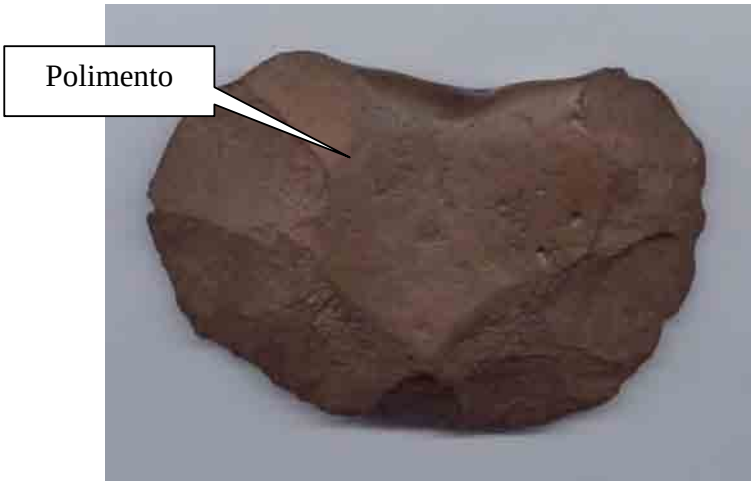


Figura 32: Rocha metamórfica proveniente do processo de silicificação

Em resumo o modelo apontou as áreas de baixa, média e alta probabilidade de ocorrência arqueológica, cuja localização dos pontos, já conhecidos, coincidem com as áreas de alta probabilidade, ou seja, não há observações que fogem do padrão conhecido (*outlier*), como mostra a figura 23.

A áreas de influências (*buffers*) se destacam na validação de que a proximidade das fontes d'água é um dos fatores determinantes para os padrões de estabelecimento de locais arqueológicos, bem como a altimetria, que aponta a presença dos vestígios nos locais mais elevados, sendo verificado no mapa de predição.

O modelo de regressão logística mostrou-se adequado para a predição de locais arqueológicos, pois esse tipo de dado é dicotômico (ausência ou presença de um determinado

evento), assim sendo o modelo é de grande valia, pois permite predizer um determinado evento dicotômico baseado em variáveis dependentes do tipo nominal, escalar, intervalar ou taxa.

O conhecimento da população que habitava a região foi de suma importância, pois os modelos preditivos só puderam ser aplicados corretamente, pois existiam padrões ambientais e padrões culturais dos habitantes na área de estudo.

Caso não existissem tais padrões, os modelos preditivos poderiam ter erros inerentes à inapropriada modelagem dos dados.

Com base no conhecimento da população pretérita que habitava a região e, apoiado na pesquisa de Warren (1990), foi determinado um modelo do tipo logito que melhor representa a probabilidade da ocorrência de sítios arqueológicos. Foi possível também predizer locais de possíveis ocupações pretéritas nas áreas não prospectadas pelos arqueólogos, utilizando modelo de regressão logística para dados arqueológicos, baseado em variável dependente dicotômica que foi ausência ou presença de sítios arqueológicos e em variáveis independentes, que explicaram o fenômeno da variável dependente, tais como as características da área de pesquisa.

7.1. Recomendações

Como os dados utilizados nesse trabalho eram de uma pesquisa do tipo de "Contrato", as informações coletadas para o mesmo não estavam preparadas para aplicação do modelo de regressão logística. Assim sendo, recomenda-se, na elaboração de um projeto de salvamento, a coleta de dados visando à aplicação de um modelo logístico, podendo coletar uma variedade maior de pontos entre áreas separadas em toda extensão e localizar todo material cartográfico anterior às etapas do Projeto de Salvamento, no qual algumas dessas amostras seriam excluídas na criação do modelo, para serem utilizadas como teste de validação do modelo.

Infelizmente, não foi possível validar o método aplicado, pois a inundação da barragem foi realizada e, seria necessário gerar uma extensão maior da área de pesquisa. Entretanto, isto demanda um intervalo de tempo muito maior do que o disponível para o presente projeto.

As idéias apresentadas neste trabalho podem ser estendidas a outros projetos, recomendando-se realizar um levantamento em várias regiões da área total, utilizando-se parte dessas amostras para gerar o modelo e a outra parte para a validação do modelo. Após isso, o modelo pode ser usado para uma outra área semelhante a essa ocupação.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, C. C. F. - Álgebra de mapas e suas aplicações em Sensoriamento Remoto e Geoprocessamento, Dissertação de Mestrado em Sensoriamento Remoto, INPE, São José dos Campos, 1997.

BORGES, K. A. V. – Modelagem de dados geográficos: Uma extensão para aplicações geográficas, Belo Horizonte: Fundação João Pinheiro, 1999.

BONA, L. D.; LARCOMBE L. - Evaluating Archaeological Potential. Artigo publicado no Quebec Ministry of Cultural Affairs, Quebec City - EUA, 1992.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J., Classification and Regression Trees, Pacific Grove: Wadswork, 1984.

CALDARELLI, S. B. - Arqueologia de Contrato no Brasil, Revista USP/ Coordenadoria de Comunicação Social, Universidade de São Paulo – N1, São Paulo, 1989.

CÂMARA, G. e MEDEIROS, J. S. de. - Geoprocessamento para Projetos Ambientais. INPE, São José dos Campos, São Paulo, 2ª ed. 1996.

CÂMARA, G.; CASANOVA, M. A.; HERMERLY A. S.; MAGALHÃES G. C.; MEDEIROS, C. M. B. – Anatomia de Sistemas de Informação Geográfica, Campinas – Instituto de Computação, Unicamp, 2002.

CAMPBELL, H.; MASSER I. – GIS and Organizations: How effective are GIS in pratice, Taylor & Francis, London, 1995.

CARVALHO, L. A. V. – *Datamining:* A mineração de dados no Marketing; Medicina, Economia, Engenharia e Administração, Editora Érica, - São Paulo 2001.

CORDEIRO, J.P.S.; AMARAL, S.; FREITAS, U.M.; CÂMARA, G. - Álgebra de Geo-Campos e Suas Aplicações. In: Simpósio Brasileiro de Sensoriamento Remoto, 1996. *Anai*s, São José dos Campos, INPE, 1996

CORDEIRO, G. – Introdução à Teoria da Verossimilhança – 10º Simpósio Nacional de Probabilidade – UFRJ 1992.

COX, D. R.; HINKLEY D. V. Monographs on Statistics and Applied Probability: Analysis of Binary Data – Chapman and Hall, London –1969.

DAVIS, C.; PAIVA, J. A.; CASANOVA, M. A. e CÂMARA, G. – Banco de Dados Geográficos. http://www.dpi.inpe.br/gilberto/livro/bdados/index.htm - 2002.

DAVIS JR., C., BORGES, K. A. V. – OMT-G: Resumo. Belo Horizonte: UFMG/IGC, 2000.

ENVIRONMENTAL SYSTEM RESEARCH INSTITUTE - Understanding GIS: The Arc/Info Method, ESRI, United Kingdom, 1995.

ENVIRONMENTAL SYSTEM RESEARCH INSTITUTE - Arc/Info Data Management: Concepts data models, database design, and storage, ESRI, United Kingdom, 1994.

GALO, M. L. T. – Notas de aulas da disciplina de Sensoriamento Remoto II – Unesp, Presidente Prudente – SP, 2000.

HOSMER, D. W.; LEMESHOW JR., S. - Applied logistic regression, John Wiley & Sons, New York, 1989.

JOHNSON, R. A.; WICHERN, D. W. – Applied Multivariate Statistical Analysis (3rd. E.), Upper Saddle River, N.J.: Prentice Hall, 1992

KVAMME, K.L. The fundamental principles and practice of predictive modelling. In A. Voorrips (ed) 1990. Mathematics and Information Science in Archaeology: a Flexible Framework: 257-295. Bonn: Studies in Modern Archaeology 3, Holos-Verlag 1990.

KVAMME, K. L. - A view from across the water: the North American esperience in archaeological GIS. En: Lock, G. and Stancic, Z. (eds.) Archaeology and Geographical Information Systems: A European Perspective. Taylor & Francis, London - 1995.

LAUDON, K. C.; LAUDON P. L. Sistemas de Informação e a Internet – LTC, Rio de Janeiro, 1999.

LIMP, W. F. e CARR, C. – The analysis of decision making: alternative applications in archaeology. In For concordance in Archaeological analysis: Bridging Data Structure, Quantitative Technique, and Theory, edited by C. Carr, Kansas City, p 128-172,1985.

MARDI, A. K. V.; KENT, J. T.; BIBBY, J. M. – Multivariate Analysis, Londres, Academic, 1992.

MEYER, P. L. – Probabilidade: aplicações à estatística; tradução do Prof. Ruy de C. B. Lourenço Filho. Rio de Janeiro, Livros Técnicos e Científicos, 1978.

MORAIS, J. L. – Arqueologia de salvamento no Estado de São Paulo, Dédalo Revista Anual de Arqueologia e Etnologia, n.º28, São Paulo – 1990.

PEREIRA, J. M. C.e ITAMI, R. M., GIS – Based Habitat Modelling using logistic multiple regression: A study of the Mt. Grahan Red Squirre, *Photogrammetric Engineering & Remote Sensing*, Vol. 57, nº. 11, pp. 1475-1486, Novembro 1991.

QUINLAN, J. R. - C4.5: Programs for machine learning, Morgan Kaufmann, Los Altos, 1993.

RUMBAUGH, J.,BLAHA, M., PREMERLANI, W., EDDY, K. AND LORENSEN, W. - Object-Oriented Modeling and Design, Prentice Hall, Englewood Cliffs, N.J - 1991.

TACHIBANA, V. M. – Apostila do PPGCC: Análise Multivarida, Unesp – Presidente Prudente – SP, 2000.

TACHIBANA, V. M. – Notas de Aula do Curso de Estatística: Regressão Logística, Unesp – Presidente Prudente – SP, 2000.

THOMAZ, R. C. C. – O uso de SIG na predição da localização de sítios arqueológicos: um estudo de Caso na bacia do Paraná superior", Tese de Doutorado, São Paulo, FFLCH-USP – 2002.

TEIXEIRA, A. L.A & MORETTI, E. e CHISTOFOLETTI, A -. Introdução aos Sistemas de Informação Geográfica, Editora do Autor. Rio Claro, São Paulo. 1997.

WARREN, R.E. - Predictive modelling of archaeological site location: a primer. In K.M.S. Allen, S.W. Green, and E.B.W. Zubrow, (eds) Interpreting Space: GIS and archaeology: 90-111. London: Taylor & Francis, 1990.

WORBOYS, M. F.- GIS – A Computing Perspective – Taylor & Francis, London, 1995.

BIBLIOGRAFIA CONSULTADA

BURROUGH, P. A. – Principles of Geographical Information Systems for Land Resources Assessment, Clarendon Press, Oxford, 1986.

BUSSAB, W. O. – Estatística Básica (3ª ed.), São Paulo, Atual, 1986.

DOBSON, A. J. – An introduction to generalized linear models – Chapman & Hall, London, 1997.

DRAPER, N. R.; SMITH H. Applied regression analysis – John Wiley & Sons, New York, 1981.

FONSECA, J. S.; MARTINS, G. A.; TOLEDO G. L. – Estatística aplicada – Editora Atlas, São Paulo – 1985.

FONSECA, J. S.; MARTINS, G. A.- Curso de Estatística – Editora Atlas, São Paulo – 1980.

FOX, J. – Linear statistical models and related methods, John Wiley & Sons, Inc. – Toronto, 1984.

FRANK, A. U.; CAMPARI I. (Eds.) – Spatial Information Theory: A theorical Basis for GIS, Springer-Verlag, Berlin, 1993.

GEMAEL, C. Introdução ao ajustamento de observações – Aplicações geodésicas, UFPR – Curitiba, 1996.

GOODCHILD, M. - A spatial analytical perspective on geographical information systems. International Journal of Geographical Information Systems. New York: Oxford University Press, 1 (4): 327-334, 1987.

GOODCHILD, M.; BRADLEY, P.; STEYAERT, L. - Environmental Modelling with GIS. New York: Oxford University Press, 1993.

GOODCHILD, M.; MAGUIRE, D. J.; RHIND, D. -Geographical Information Systems: Principles and applications. (2 volumes) New York: John Wiley and Sons, 1991.

GUIDORIZZI, H. L. Um curso de Cálculo Vol. 1 – 2ª ed. – Rio de Janeiro; São Paulo: LTC – Livros Técnicos e Científicos Editora S/a, 1987.

GUIDORIZZI, H. L. Um curso de Cálculo Vol. 2 – 3ª ed. – Rio de Janeiro; São Paulo: LTC – Livros Técnicos e Científicos Editora S/a, 1987.

GOMES, J.; VELHO, L. – Computação Gráfica: Imagem – Rio de Janeiro – IMPA, 1994.

KARSON, M. J. – *Multivariate statistical methods*, Iowa State University – Iowa – 1982.

KASHIMOTO, E. M. - Variáveis ambientais e arqueologia no Alto Paraná. Tese de Doutorado, São Paulo, FFLCH-USP – 1997.

KOCK JR., N. F., MCQUEEN, R. J. AND FERNANDES, C. T., "Information Systems Research in Organizations: An Action Research Approach", Brazilian Journal of Contemporary Management, V.1, No.4, 1995, pp. 155-175.

KOHLER, T.A. e PARKER, S.C.. Predictive models for archaeological resource location. In M. B. Schiffer (ed.) Advances in Archaeological Method and Theory, Vol. 9: 397-452. New York: Academic Press, 1986.

LOCK, G.; STANCIC, Z. – Archaeology and geographical information systems – Talylor & Francis, London, 1995.

LUCAS JR, H.C. - Information Technology for Management, McGraw-Hill, 1997.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. – Multivariate Analysis, Londres, Academic 1992.

MOALA, F. A. Notas de Aula Probabilidade e Estatística, Unesp – Presidente Prudente, 1998.

MOSTELLER, F.; TUKEY J. W - Data analysis and regression: a second course in statitics, Addison – Wesley Publishing Company, Massachusetts – 1977.

RAMINELLI, J. A.; TACHIBANA, V. M. – Introdução às regressões linear e logística: Relatório final de Estágio Obrigatório apresentado para a Disciplina de Estatística Aplicada, Unesp - Presidente Prudente, 2000.

RUSTAGI, J. S. – Introduction to Statistical Methods (Vol. 1) – New Delhi, Rowman & Allanheld, 1984.

SEBER, G. A. F. – Linear regression analysis – John Wiley & Sons, - New York, 1997.

SETZER, V. W. - Meios Eletrônicos e Educação: uma visão alternativa_ Coleção Ensaios Transversais No. 10. São Paulo: Ed. Escrituras, p. 239-275.

SHENNAN, S. – Quantifying Archaeology – Edinburgh University Press, United Kingdown 1988.

UPTON, G. J. C. – Spatial data analysis by example, Vol.1 John Wiley & Sons Ltd. New York, 1988.

SITES DE REFERÊNCIA

http://www.prudente.unesp.br/dcartog.htm
http://www.ime.usp.br/~vwsetzer
http://www.dpi.inpe.br/
http://www.esri.com/library/userconf/proc96/TO150/PAP119/P119.HTM
http://www.cis.temple.edu/~kock/public/tlo97/kiflww6.htm
http://www.fao.org/docrep/W5830E/w5830e0k.htm#basic concepts
http://www.iphan.gov.br/legislac/legisla.htm
http://www.cast.uark.edu/~kkvamme/sieber/sieber.htm
http://www.cast.uark.edu/~kkvamme/mnmodel/mnmodel.htm
http://www.lei.adv.br/001-86.htm