CENTRO DE ESTATÍSTICA APLICADA – CEA - USP

RELATÓRIO DE CONSULTA

TÍTULO**:** Relatório de consulta sobre o projeto “Análise dos padrões espaço-temporais de distribuição do boto cinza *Sotalia fluviatilis* na região de Cananéia”

PESQUISADOR**:** Mario Manoel Rollo Junior.

ORIENTADOR**:** Mario de Vivo.

INSTITUIÇÃO**:** Museu de Zoologia da Universidade de São Paulo

FINALIDADE DO PROJETO**:** Doutorado

PARTICIPANTES DA ENTREVISTA: Mario Manoel Rollo Junior
Mario de Vivo
Júlia Maria Pavan Soler
Carlos Alberto de Bragança Pereira
Edgard Rodrigues Fusaro

DATA: 23/05/2000

FINALIDADE DA CONSULTA: Sugestões para levantamento e análise de dados.

RELATÓRIO ELABORADO POR: Edgard Rodrigues Fusaro

1. INTRODUÇÃO

Os cetáceos, que são mamíferos marinhos, apresentam uma admirável relação de harmonia quando se relacionam com outros componentes de uma certa fauna local. Em ambientes estuarinos e marinhos, estes animais são possíveis indicadores-chave de qualidade ambiental. Dessa forma, modelos preditivos de distribuição e abundância podem ser elaborados a partir do uso destes mamíferos.

Durante um perído de dois anos, foram observadas, em todo o contorno da Ilha de Cananéia (litoral do Estado de São Paulo), a abundância (densidade) de botos, a temperatura superficial da água e a salinidade na superfície. Assim, segue que o objetivo do estudo é combinar o uso de Modelos Estatísticos e Sistemas de Informação Geográfica (SIG's) na elaboração de modelos que possibilitem prever a distribuição destes animais na região ao longo do tempo. A idéia seria utilizar estes modelos em áreas com características ambientes semelhantes, que também são habitadas pelos botos, como forma de previsão de abundância e densidade.

2. DESCRIÇÃO DO ESTUDO E DO PROCESSO DA COLETA DE DADOS

A região onde está situada a Ilha de Cananéia oferece uma infra-estrutura para a realização deste tipo de trabalho. O impacto de atividades humanas, como por exemplo esgotos e lixo, é bem baixo nesta região. Os botos movimentam-se por toda a extensão da ilha, algo que não foi avaliado antes e que faz com que o presente trabalho seja uma revisão de trabalhos feitos na região de Cananéia.

Primariamente, os animais não estão na área por motivo de estarem fazendo uso de alimentos existentes na mesma. A agregação dos animais em determinadas áreas da ilha influencia na densidade. Dessa forma, todo o contorno da ilha foi subdivido em 19 áreas de estudo. A observação dos animais foi realizada a partir de transectos lineares utilizados em cada uma das áreas de

amostragem. Aleatoriamente, uma área era selecionada e determinada como o ponto de partida do trabalho de campo; em seguida, a seqüência das áreas a serem percorridas era escolhida. Cabe aqui ressaltar que, durante o trabalho de coleta de dados, não foi possível percorrer cada uma das 19 áreas em apenas um dia de campo (o máximo que se conseguiu amostrar foram 13 áreas), sendo necessários 3 ou 4 dias para que todas as áreas fossem abrangidas. Para tal, era realizado um sorteio no início do próximo dia de coleta para determinar a nova área de partida; este sorteio era feito de forma a não começar o trabalho de campo a partir de uma área que já havia sido amostrada no dia anterior.

Para cada um dos animais observados, foi obtido um ponto referente às suas coordenadas (posição geográfica); além disso, eram calculados, com referência à posição em que o animal era observado a partir do barco da equipe de campo, uma distância (medida em metros) e um ângulo (medido em graus).

Todos os dados coletados durante o período de 24 meses foram transmitidos eletronicamente para o computador. Mais precisamente, em cada um dos 19 pontos de coleta, em cada mês do trabalho de campo, foram obtidas informações sobre a densidade de botos (medida em nº ind./km^2), a temperatura superficial da água (graus Celsius), a salinidade na superfície (ppm) e variáveis categóricas como o tipo de substrato, a pluma de sedimentos, o estado de conservação de cercos de pesca e a concentração de clorofila (todas avaliadas através de análise de imagens aerofotogramétricas e de sensoriamento remoto).

3. SUGESTÃO DO CEA

3.1 Análise Descritiva

Em nosso estudo, a variável dependente (resposta) é a densidade de botos. Dessa forma, temos que as variáveis independentes do estudo são: temperatura superficial da água, salinidade na superfície, tipo de substrato, pluma de sedimentos, estado de conservação de cercos de pesca e concentração de

clorofila, todas estas associadas a um dos 19 pontos de coleta em um determinado mês.

Primeiramente, para cada uma das 19 áreas, e independente do mês de coleta, poderíamos calcular medidas descritivas como a média, o desvio padrão e a mediana para as variáveis densidade de botos, temperatura superficial da água e salinidade na superfície; para as variáveis categorizadas (tipo de substrato, pluma de sedimentos, estado de conservação de cercos de pesca e concentração de clorofila), construiríamos tabelas com a distribuição de freqüências para cada uma delas. A mesma análise das variáveis poderia ser feita para cada mês, independente da área. Assim, pode-se avaliar qual fonte de variabilidade, mês ou área, modifica mais as variáveis estudadas.

A seguir, seria interessante verificar o efeito que cada uma das variáveis contínuas exerce sobre a resposta quando tomada de forma isolada das demais, ou seja, os seguintes gráficos de dispersão bidimensionais, obtidos primeiramente para uma área fixada (independente do mês) e depois para um mês fixado (independente da área), poderiam ser construídos:

i) densidade de botos X temperatura superficial da água

ii) densidade de botos X salinidade na superfície

Neste caso, poderiam ser calculados os correspondentes coeficientes de correlação linear de Pearson para os casos “i” e “ii”.

Já em relação às variáveis categorizadas, considerando cada um dos 19 pontos de coleta e fixando um certo mês, poderiam ser construídas as seguintes tabelas de contingência de dupla entrada:

i) tipo de substrato X pluma de sedimentos

ii) tipo de substrato X estado de conservação de cercos de pesca

iii) tipo de substrato X concentração de clorofila

iv) pluma de sedimentos X estado de conservação de cercos de pesca

v) pluma de sedimentos X concentração de clorofila

vi) estado de conservação de cercos de pesca X concentração de clorofila

Dessa forma, para os dados de cada casela das tabelas acima, poderíamos calcular medidas descritivas (média, desvio padrão, mediana) da variável

densidade de botos. Detalhes sobre estes procedimentos de análise descritiva podem ser encontrados, por exemplo, em Bussab e Morettin (1987).

3.2 Análise Inferencial

A equação do modelo de regressão usual é da forma

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \varepsilon_j \ , \ j = 1, 2, ..., n$$

onde as variáveis residuais ε_j devem satisfazer as seguintes suposições (ver Neter et al.(1996), por exemplo):

i) $E(\varepsilon_j) = 0, \forall j \Rightarrow E(y_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ...$

ii) $Var(\varepsilon_j) = \sigma^2, \forall j \Rightarrow Var(y_j) = \sigma^2, \forall j$ (suposição de homocedasticidade)

iii) $\varepsilon_i, \varepsilon_j$, são variáveis aleatórias independentes, $i \neq j$ e $i, j = 1, 2, ..., n$

iv) $\varepsilon_j \sim N(0, \sigma^2) , \forall j \Rightarrow y_j \sim N(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ..., \sigma^2)$

A variável dependente y_j é expressa como uma combinação linear de fatores independentes e covariáveis.

Nos modelos loglineares (ver Agresti(1990), por exemplo), a variável a ser predita é uma contagem (que aparece no lado esquerdo da equação, como no modelo de regressão) e a equação original é exponencial, isto é,

$$m_j = \exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ...)$$

onde $m_j = E(n_j)$, com m_j representando o valor esperado da contagem e n_j representando o valor observado da contagem, sendo que, de uma forma geral, temos $n_j \sim$ Poisson (λ_j).

Quando tomamos o logaritmo natural de ambos os lados da equação, chegamos a

$$\ln(m_j) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$$

O log da contagem é expresso como uma combinação linear de fatores e covariáveis. Entretanto, é fácil converter os valores obtidos em log nos verdadeiros valores da contagem, a partir do cálculo da exponencial. Por exemplo, se o ln (logaritmo neperiano) da contagem for 1,98, teremos que o verdadeiro valor da contagem será

$$e^{1,98} = 7,21.$$

Assim, pode-se supor que a variável dependente do estudo (contagem de animais) tem uma possível distribuição de Poisson (este fato pode ser considerado razoável se obtermos, a partir dos dados desta variável, um valor para sua média aproximadamente igual ao valor para sua variância). Neste caso, pode ser sugerido para a análise o uso do modelo loglinear de Poisson. Duas suposições muito importantes que devem ser observadas antes de fazermos uso do modelo loglinear são:

- o tamanho total da amostra não é fixo antes do estudo ou a análise não é condicionada ao tamanho da amostra, isto é, não conhecemos a priori a densidade de botos nem tampouco o modelo loglinear será condicionado a esta densidade, uma vez que só depende dos valores observados para cada uma das covariáveis e fatores
- existe independência entre as observações de diferentes caselas do modelo, ou seja, a densidade de animais observadas em um determinado ponto, num certo mês, sob determinadas condições ambientais, não depende da densidade medida em um outro ponto, num certo mês, submetido a condições ambientais específicas

Portanto, segue que um possível modelo a ser utilizado é dado por:

$$\ln(y_{ij}) = \mu + \text{efeitos principais} + \text{efeitos de interação} + \text{resíduo}$$

$$i = 1, 2, ..., 19 \text{ (pontos)}, j = 1, 2, ..., 12 \text{ (meses)}$$

onde:

y_{ij} = densidade de botos no i-ésimo ponto no j-ésimo mês;

μ = média geral;

efeitos principais = efeito de cada uma das variáveis do estudo tomada de forma isolada das demais: variáveis contínuas (temperatura superficial da água, salinidade na superfície), variáveis categorizadas (tipo de substrato, pluma de sedimentos, estado de conservação dos cercos de pesca, concentração de clorofila) e os fatores mês e área;

efeitos de interação = todos os possíveis efeitos de interação das variáveis combinadas duas a duas e demais ordens mais altas;

resíduo = erro aleatório pertinente ao modelo.

É importante observar que, se o número de variáveis preditivas do estudo for muito grande teremos uma grande quantidade de parâmetros a serem estimados.

Existem, ainda, modelos que estendem o padrão do modelo de regressão linear, como, por exemplo, os modelos aditivos (ver Hastie e Tibshirani (1990), por exemplo), que assumem a média da resposta ser modelada como uma soma de variáveis preditoras. Outra classe muito útil de modelos lineares são os Modelos Lineares Generalizados (M.L.G. ’s), os quais representam uma generalização de modelos de regressão linear. Especificamente, os efeitos preditores são assumidos serem lineares nos parâmetros, mas a distribuição das respostas, bem como a *ligação* entre os preditores e sua distribuição, podem ser muito gerais. Uma extensão desta classe de M.L.G. ’s são os chamados Modelos Aditivos Generalizados (do inglês Generalized Additive Models – G.A.M. ‘s). Eles estendem os M.L.G. ‘s da mesma maneira como os modelos aditivos estendem o modelo de regressão linear, isto é, substituem a forma linear

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \varepsilon_j \ , \ j = 1, 2, ..., n$$

pela forma aditiva

$$y_j = \beta_0 + f_1(x_{1j}) + f_2(x_{2j}) + ... + \varepsilon_j \ , \ j = 1, 2, ..., n$$

Alguns detalhes sobre a estrutura destes modelos é apresentada no Apêndice.

4. CONCLUSÃO

Este relatório contém uma orientação técnica inicial para uma análise exploratória dos dados do projeto em questão. Sugere-se aos pesquisadores o retorno ao CEA assim que os dados já estiverem devidamente coletados e armazenados em planilhas (por exemplo, Microsoft Excel) e submetam o trabalho para triagem de projetos a serem analisados através do serviço do CEA.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. (1990), **Categorical Data Analysis,** John Wiley & Sons, 558 p.

BUSSAB, W.O. e MORETTIN, P. A.(1987), **Estatística Básica,** 4ª Ed. Atual Editora Ltda. São Paulo, 321 p.

HASTIE, J., TIBSHIRANI, M. (1990), **Generalized Additive Models,** Chapman & Hall. London, 335 p.

NETER, J., KUTNER, M.H., NACHTSHEIN, C.J. e WASSERMAN, W. (1996), **Applied Linear Statistical Models**, 4th. ed. Irwin.

SPSS INC. (1995), **SPSS Advanced Statistics 7.5**, 1st. ed. Chicago: SPSS Inc.

APÊNDICE

Um M.L.G. consiste de uma *componente aleatória*, uma *componente sistemática* e uma *função de ligação*, que liga as duas componentes. A resposta Y é assumida pertencer a família exponencial, cuja densidade é dada por

$$\rho_Y(y;\theta;\phi) = \exp[\ (y\theta - b(\theta))/a(\phi) + c(y,\phi)\ ] \qquad (*)$$

onde θ é chamado de parâmetro natural, e ϕ é o parâmetro de dispersão. Esta é a componente aleatória do modelo. É também assumido que o valor esperado de Y, chamado de μ, está relacionado à série de covariáveis X_1, X_2, X_3, ... através de uma função $g(\mu) = \eta$, onde $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$ η é a componente sistemática, chamada de preditor linear, e g(.) é a função de ligação. Observe que a média μ está relacionada com o parâmetro natural θ por $\mu = b'(\theta)$; uma ligação óbvia para qualquer ρ dado é chamada de *ligação canônica*, a qual admite que $g(\mu)$ é escolhida tal que $\eta = \theta$. Por exemplo, para a distribuição de Bernoulli, temos que E(Y) = P(Y = 1) é a média e as ligações geralmente usadas são *logito* e *probito*, com a *logito* sendo a ligação canônica. É de costume, entretanto, definir o modelo em termos de μ e $g(\mu) = \eta$ e, então, θ não desempenha o papel de ligação. Assim, quando for conveniente, escrevemos $\rho_Y(y;\theta;\phi)$ como sendo $\rho_Y(y;\mu;\phi)$. A estimação de μ não envolve o parâmetro de dispersão ϕ, pois como forma de simplicidade ele é assumido conhecido.

Dada uma escolha específica das componentes aleatórias e sistemáticas, além de serem determinados a função de ligação, o vetor de *n* observações de y, e *p* correspondentes vetores preditores x_1, x_2, ..., x_p, o estimador de máxima verossimilhança de $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ é definido pelas equações escore:

$$\sum x_{ij} (d\mu_i/d\eta_i)\ V_i^{-1}\ (y_i - \mu_i) = 0, \quad j = 0,1, ..., p$$

onde $V_i = var\ (Y_i)$. Note que, para a equação acima, assume-se que $x_{i0} = 1$.

Um G.A.M. difere de um M.L.G. no fato de que um preditor *aditivo* substitui o preditor *linear*. Especificamente, assume-se que a resposta Y tem uma distribuição dada pela equação (*), com a média $\mu = E(Y/X_1 , ..., X_p)$ ligada aos preditores através de

$$g(\mu) = \beta_0 + f_1 (x_1) + f_2 (x_2) + ... + f_p (x_p)$$

A estimação de β_0 e f_1, f_2, ..., f_p é realizada substituindo-se os pesos da regressão linear obtidos pelo ajuste das variáveis dependentes do modelo de regressão linear por um algoritmo apropriado que ajusta os pesos referentes ao modelo aditivo.