UNIVERSIDADE FEDERAL DE PERNAMBUCO

RELATÓRIO DE ESTÁGIO SUPERVISIONADO

BACHARELADO EM ESTATÍSTICA

APLICAÇÃO DA REGRESSÃO GAMA NO MODELO POLINOMIAL PARA EXPLICAÇÃO DA CONCENTRAÇÃO DE FÓSFORO TOTAL PRESENTE NO ESTUÁRIO DO RIO GOIANA PE/PB

Edijan Cavalcanti de Assis Júnior

Orientadora Dra. Claudia Regina Oliveira de P. Lima

Co-orientador Dr. Mario Barletta

RECIFE

2012

EDIJAN CAVALCANTI DE ASSIS JÚNIOR

APLICAÇÃO DA REGRESSÃO GAMA NO MODELO POLINOMIAL PARA EXPLICAÇÃO DA CONCENTRAÇÃO DE FÓSFORO TOTAL PRESENTE NO ESTUÁRIO DO RIO GOIANA PE/PB

Relatório de estágio supervisionado apresentado ao departamento de estatística da Universidade Federal de Pernambuco, como parte dos requisitos para obtenção do grau de bacharel em estatística.

Orientadora Dra. Claudia Regina Oliveira de P. Lima
Co-orientador Dr. Mario Barletta

RECIFE
2012

UNIVERSIDADE FEDERAL DE PERNAMBUCO

BACHARELADO EM ESTATÍSTICA

Este relatório foi julgado adequado para obtenção da aprovação na disciplina Estágio Supervisionado do curso de Bacharelado em Estatística da Universidade Federal de Pernambuco – UFPE.

Professora Claudia Regina Oliveira P. Lima

Orientadora

Professor Mario Barletta

Co-orientador

Professora Audrey Helen Mariz de Aquino Cysneiros

Coordenadora do curso

RECIFE

2012

AGRADECIMENTOS

Ao meu Senhor, que colocou muitas experiências e pessoas magníficas na minha vida.

Aos meus Pais, Edijan Cavalcanti e Germana Araújo, que ofereceram tudo de bom na minha vida e me auxiliaram nos momentos difíceis com bastante sabedoria.

A minha futura esposa, Bruna Eloize, pelo o amor, compreensão, carinho, sabedoria, sorte e por proporcionar momentos felizes e marcantes em minha vida.

A minha irmã, Nyara Cavalcanti, que tenho como filha, pela amizade e respeito que têm comigo.

A professora, Claudia Regina que ajudou constantemente no desenvolvimento desse trabalho.

Ao professor Mario Barletta por me ensinar a filosofia da vida e por me guiar e contribuir – indiscutivelmente - no meu desenvolvimento acadêmico.

A professora, Mônica Costa, por auxiliar na parte filosófica e teórica desse trabalho.

Aos meus amigos, Daniel Cassimiro, Rodrigo Almeida, Natalia Pires e Williames Hora, por me ajudar nas constantes dúvidas e por proporcionar momentos felizes ao longo do curso.

SUMÁRIO

1. INTRODUÇÃO 1

2. VARIÁVEIS DE ESTUDO 2

3. MODELO LINEAR GENERALIZADO (MLG) 4

3.1. Definição 5

3.2. Distribuição Gama 6

3.3. Função de Ligação 6

3.4. Estimação dos Parâmetros 7

4. MODELO POLINOMIAL 8

5. RESULTADOS 8

6. CONCLUSÃO 15

1. INTRODUÇÃO

Os estuários são um dos ecossistemas mais influenciados pelo aporte de águas terrestres, pois como a própria etimologia diz, estuário é um termo advindo da palavra latina *aestuarium*, a qual significa pântano ou canal, que por sua vez, é derivada da palavra *aestum*, que significa maré, ou movimento de ondas (Elliot & McLusky 2002). Este termo é utilizado genericamente para indicar o encontro do rio com o mar, caracterizando uma foz litorânea. Portanto, trata-se de um ecossitema de transição entre o oceano e o continente; a complexibilidade e a vulnerabilidade à influência do homem são características comuns a todos os estuários. Em condições naturais, os estuários são biologicamente mais produtivos do que rios e o oceano adjacente, por apresentarem altas concentrações de nutrientes que estimulam a produção primária (Miranda *et al*. 2002).

Estuários são ambientes de alta riqueza de espécies, biomassa abundante, e diversidade de fatores biológicos (reprodução, desova e recrutamento) e processos ecológicos (conectividade entre o continente e a região costeira adjacente e aporte de nutrientes) (Barletta & Costa 2009). Portanto, em resultado de sua produtividade e recursos prontamente disponíveis, estuários têm sido sempre diretamente responsáveis pela subsistência das populações tradicionais em todo o mundo (Barletta & Costa 2009).

Transportado através das águas dos rios para os oceanos e acumulando-se em parte nas águas estuarinas, o elemento químico fósforo é assimilado por bactérias, plantas bênticas e remineralizado por atividades heterotróficas dos animais e micro-organismos (Santos, 2007). O fósforo também é assimilado pelo fitoplâncton que é considerado o principal produtor primário dos ambientes costeiros, sendo responsável pelo inicio do fluxo de matéria e energia da rede trófica destes ambientes, contribuindo para a sua fertilização, sustentando diretamente os herbívoros e indiretamente os animais dos níveis tróficos superiores, incluindo espécies economicamente importantes (Dring 1992).

A área estuarina do rio Goiana está situada no extremo norte do estado de Pernambuco, na divisa com o Estado da Paraíba. O rio Goiana tem sua origem após o encontro dos rios Capibaribe Mifim com o Tracunhaém, a quatro quilômetros à jusante da cidade de Goiana. O rio Goiana possui 17,5 km de extensão desde as imediações da cidade até a sua foz. Esta área estuarina corresponde a 4.776 hectares, sendo uma das maiores reservas biológicas da região Metropolitana do Recife (FIDEM 1987).

É relevante, portanto, a análise e modelagem das observações referentes à concentração total do fósforo (P), motivadas pela sua colaboração na produtividade em estuários para conhecer como esse processo ocorre. Para tanto, será aplicada uma regressão polinomial (Weisberg, 2005), a fim de evidenciar se as interações de fatores físico-químicos como a salinidade com a demanda bioquímica de oxigênio, temperatura com pH, entre outros assumem papel importante para a descrição da concentração do fósforo total no estuário do rio Goiana e região costeira adjacente.

2. VARIÁVEIS DE ESTUDO

As variáveis em estudo foram investigadas parte por meio de dispositivos eletrônicos e parte mensurada por amostras recolhidas. Estas foram capturadas em média a cada dois meses e 15 dias nos períodos 1993-1996 e 1999-2009 no estuário do rio Goiana PE/PB (CPRH – Agência Estadual de Meio Ambiente). Os fatores abióticos foram capturados em três estações do rio Goiana, as estações GO-67, GO-80 e GO-85 (Figura 1).

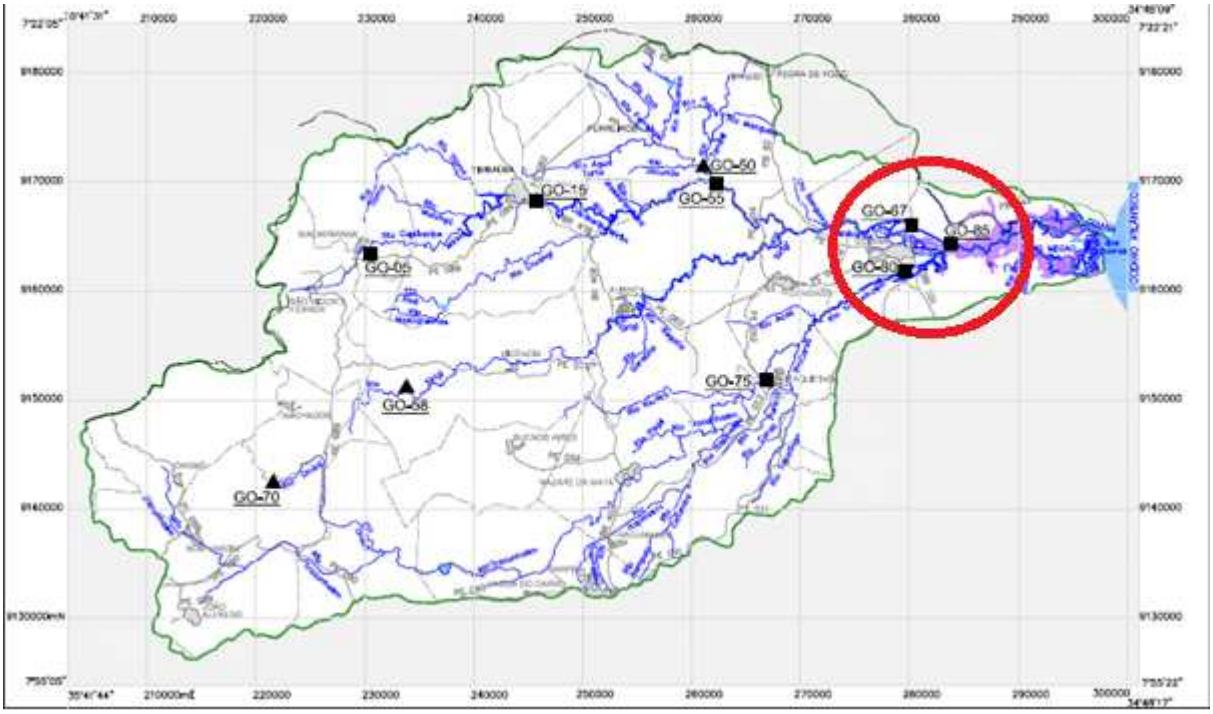

Figura 1 – Localização das estações do rio Goiana onde as informações foram capturadas nos períodos de 1993-1996 e 1999-2009, no estuário do Rio Goiana PE/PB.

Nesta parte do trabalho será efetuada uma análise descritiva acerca dos fatores físico-químicos coletados no estuário do rio Goiana e região adjacente. Na Tabela 1, a seguir, estão dispostos os valores do mínimo e máximo, média, mediana, desvio padrão e coeficiente de variação dos fatores físico-químicos.

As variáveis ambientais que serão trabalhadas nesse estudo:

- **Índice Pluviométrico:** medição em milímetros (mm) da soma mensal da precipitação de chuva no estuário do rio Goiana e região adjacente nas estações GO-67, GO-80 e GO-85;
- **Salinidade:** medição em (‰) de quantidade de sais existente em massa de águas naturais;
- **pH:** é o potencial hidrogeniônico, que indica a acidez, neutralidade ou alcalinidade no estuário do rio Goiana e região adjacente;
- **Demanda Bioquímica de Oxigênio** (**D.B.O.):** corresponde a quantidade de oxigênio consumido na degradação da matéria orgânica por processos biológicos, medida em (mg/L);
- **Temperatura:** é a temperatura medida em graus Celsius (ºC) nas estações GO-67, GO-80 e GO-85 do rio Goiana e região adjacente;
- **Concentração de coliforme fecais:** medida em NMP/100 ml coletado nas estações GO-67, GO-80 e GO-85.
- **Concentração total do fósforo (P):** é o conjunto das várias formas de fosfatos medido em (mg/L).

Tabela 1 – Medidas descritivas dos fatores físico-químicos no período de 1993-1996 e 1999-2009 no estuário do rio Goiana e região adjacente

Observação	Valor mínimo	Média	Mediana	Valor máximo	Desvio padrão	Coeficiente de variação
Pluviometria	0,14	155,81	114	344	94,2	60,5%
Salinidade	0,13	0,68	0,27	6,4	1,21	56,2%
pH	6,35	7,23	7,23	8,2	0,27	3,7%
D.B.O.	0,5	4,83	2,44	70	9,23	191,1%
Temperatura	23	28,15	28,25	31,5	1,26	4,5%
C. de coliforme fecais	200	21578	3350	160000	41866,6	194,0%
Fósforo	0,06	0,29	0,19	1,72	0,28	96,3%

Os coeficientes de variação das variáveis concentração de coliforme e D.B.O. sobressaem-se largamente aos demais (Tabela 1).

A diferença entre a média e mediana da variável fósforo aponta assimetria dos dados, ainda mais explícita no histograma a seguir apresentado (Figura 2), disposto em auxílio à visualização de possível distribuição a descrever de modo adequado a concentração total do fósforo. Distribuição Gama (Paula, 2010) é cogitada. A confirmação deu-se com a aplicação do teste de Kolmogorov-Smirnov (Ross, 2006), cuja hipótese nula de que a frequência da variável Fósforo pode ser descrita por tal distribuição, não foi rejeitada ao nível de significância de 5% (p-valor= 0,42).

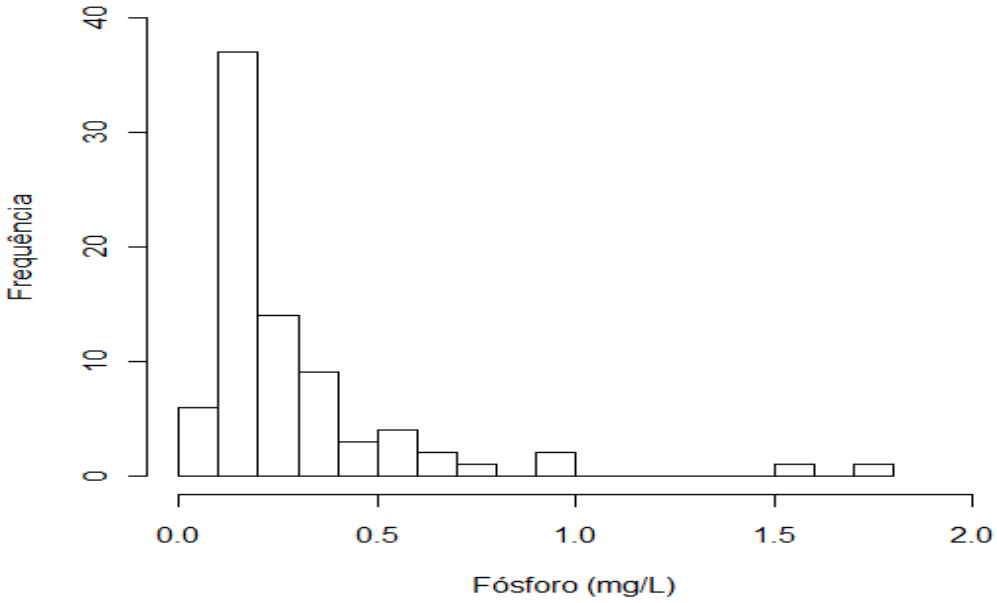

Figura 2 - Visualização da distribuição da variável resposta.

3. MODELO LINEAR GENERALIZADO (MLG)

Na década de 70, o desenvolvimento computacional permitiu trabalhar com alguns modelos que exigiam a utilização de métodos iterativos para a estimação dos parâmetros. Nelder & Weddeburn (1972) apresentaram uma proposta bastante inovadora, os modelos lineares generalizados (MLG).

A ideia básica consiste em abrir um leque de opções para a distribuição da variável resposta, permitindo assim que a mesma pertença à família exponencial e possua maior

flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear η (Paula, 2010).

3.1 Definição

O modelo linear generalizado envolve uma única variável resposta, variáveis explicativas e uma amostra aleatória de *n* observações, em que:

(i) O componente aleatório do modelo é a variável resposta, que pertence à família exponencial (Equação 2) (Demétrio 2002).

(ii) O componente sistemático é a combinação linear das variáveis explicativas (Demétrio 2002).

(iii) A ligação entre o componente aleatório e o componente sistemático é realizada através de uma função de ligação g(.), monótona e duas vezes diferenciável (Demétrio 2002).

$$f(y_i;\theta_i;\phi) = \exp[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)], \forall i. \tag{1}$$

A média da variável resposta é dada por

$$E[Y_i] = b'(\theta_i) = \mu_i, \tag{2}$$

e a variância por

$$Var(Y_i) = \phi^{-1}V(\mu_i), \tag{3}$$

$V_i = V(\mu_i) = \dfrac{d\mu_i}{d\theta_i}$ sendo função de variância, e $\phi^{-1} > 0$ parâmetro de dispersão. A função de variância representa o fator da componente da variância da variável resposta Y que depende da média μ. Com a função de variância tem-se uma classe de distribuição correspondente e vice-versa (Paula, 2010).

A parte sistemática do modelo é definida por:

$$g(\mu_i) = \eta_i, \tag{4}$$

em que, $\eta_i = \mathbf{x}_i^T\beta$ é o preditor linear, $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$, $p < n$, é um vetor de parâmetros desconhecidos a serem estimados, $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})^T$ representa os valores das variáveis explicativas e $g(.)$ função de ligação.

3.2 Distribuição Gama

Seja Y uma variável aleatória com distribuição Gama de média θ e coeficiente de variação $\phi^{-(1/2)}$, cuja função densidade é expressa na forma:

$$f(y;\mu;\phi) = \frac{1}{\Gamma(\phi)}\left(\frac{\phi y}{\mu}\right)^{\phi} \exp\left(-\frac{\phi y}{\mu}\right) d(\log y)$$

$$= \exp\left[\phi\left\{\left(-\frac{y}{\mu}\right) - \log\mu\right\} - \log\Gamma(\phi) + \phi\log\phi y - \log y\right], \qquad (5)$$

em que , $y > 0, \phi > 0, \mu > 0$ e $\Gamma(\phi) = \int_0^{\infty} t^{\phi-1}\exp(-t)dt.$

Pela Equação (5), pode-se dizer que,

$$\theta = -\frac{1}{\mu} \therefore b(\theta) = -\ln(-\theta).$$

Realizando a primeira e segunda derivada de $b(\theta)$ para obter a média e a função de variância respectivamente obtêm-se,

$$b(\theta) = -\ln(-\theta) \therefore b'(\theta) = -\frac{1}{\theta} \therefore b''(\theta) = \frac{1}{\theta^2}.$$

3.3 Função de Ligação

Supondo que o coeficiente de variação é conhecido, pode-se expressar a função de verossimilhança de um MLG com respostas independentes por

$$L(\beta) = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi). \qquad (6)$$

Quando $\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, a Equação (6) pode ser reescrita da seguinte forma

$$L(\beta) = \sum_{i=1}^{n} \phi\left\{y_i \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\right\} + \sum_{i=1}^{n} c(y_i, \phi). \qquad (7)$$

Definindo a estatística $S_j = \phi\sum_{i=1}^{n} Y_i x_{ij}$, a Equação (7) pode ser escrita por

$$L(\beta) = \sum_{j=1}^{p} s_j\beta_j - \phi\sum_{i=1}^{n} b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) + \sum_{i=1}^{n} c(y_i, \phi). \qquad (8)$$

Pelo teorema da fatorização a estatística $\mathbf{S} = (S_1, S_2, \ldots, S_p)^T$ é suficiente minimal para o vetor $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ (Paula 2010). Estas estatísticas são correspondidas pela função de ligação canônica.

Neste relatório serão abordadas três funções de ligação fornecidas na Tabela 2.

Tabela 2 – Funções de ligação que serão abordadas no modelo Gama

Função de ligação	Inversa (Canônica)	Logarítmica	Identidade
Expressão	$\eta = \mu^{-1}$	$\eta = \log(\mu)$	$\eta = \mu$

As funções de ligações canônicas conduzem a propriedades estatísticas desejáveis para o modelo, tais como, suficiência, facilidade nos cálculos, interpretação simples, entre outros. Nem sempre a escolha da função de ligação canônica pode-se esperar os melhores resultados, o aspecto importante a ter em consideração na escolha da ligação é a adaptabilidade e adequabilidade do modelo.

3.4 Estimação dos Parâmetros

Será realizado o processo iterativo de mínimos quadrados reponderados para obter as estimativas do vetor $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$. Este processo é expresso da seguinte forma

$$\beta^{(m+1)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(\mathbf{m})}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(m)}\mathbf{z}^{(\mathbf{m})}, m = 0,1,\ldots, \tag{9}$$

em que, $\mathbf{z} = \eta + \mathbf{W}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}(\mathbf{y} - \mu)$ é uma covariada modificada, enquanto $\mathbf{W}$ é uma matriz de pesos que é alterada a cada iteração. Em geral a Equação (9) converge em número de passos finito independente dos valores iniciais.

No caso da distribuição Gama, o coeficiente de variação é estimado pelo método de máxima verossimilhança, de modo particular, na seguinte equação:

$$2n\left\{\log\hat{\phi} - \upsilon(\hat{\phi})\right\} = D(\mathbf{y};\hat{\mu}), \tag{10}$$

onde $D(\mathbf{y};\hat{\mu}) = 2\sum_{i=1}^{n}\left\{-\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}\right\}$ é a função desvio e $\upsilon(\hat{\phi}) = \frac{\Gamma'(\hat{\phi})}{\Gamma(\hat{\phi})}$ é a função digama.

4. MODELO POLINOMIAL

Levando em consideração que a natureza é um conjunto de ações que ocorrem ao mesmo tempo, será abordado nesse trabalho um caso particular da regressão múltipla, a regressão polinomial (Weisberg, 2005). O ajuste arbitrário de um polinômio de grande ordem, quando o grau é maior que dois, é um sério abuso na análise de regressão (Montgomery, 2006). Um modelo de baixa ordem em uma variável transformada é quase sempre preferível a um modelo de alta ordem na métrica original (Montgomery, 2006). O modelo polinomial será efetivado com princípios hierárquico que é definido quando contém todos os termos de graus inferiores ao maior grau do modelo. A escolha de um modelo hierárquico será adotada por ser invariante sob uma transformação linear (Montgomery, 2006). Devido aos critérios citados anteriormente será utilizado o polinômio de grau dois nesse relatório.

Com mais de um preditor, obtêm-se expoentes inteiros e produto de todos os preditores como termos da função da média. Como exemplo será citado um caso de dois preditores, a função média é dada por:

$$E(\mathbf{Y} \mid \mathbf{X}) = \beta_0 + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_1^2 + \beta_4 \mathbf{x}_2^2 + \beta_5 \mathbf{x_1 x_2}. \qquad (11)$$

O novo termo $\mathbf{x_1 x_2}$ na Equação (11) é conhecido como interação. Com k preditores, o modelo de segunda ordem inclui um intercepto, *k* termos lineares, *k* termos quadráticos e $k(k-1)/2$ interações.

5. RESULTADOS

Deseja-se obter um modelo que melhor explique a concentração total do fósforo a partir dos fatores físico-químico existente nesse estuário localizado na região tropical. Os valores da concentração total do fósforo são maiores que zero com média amostral aproximadamente 0,29 e desvio padrão amostral aproximadamente a 0,28.

Além disso, a diferença entre a média e mediana da variável fósforo aponta assimetria dos dados (Figura 2) indicando, em conjunto com teste de adequacidade de Kolmogorov-Smirnov, ser plausível a utilização da distribuição gama na modelagem da variável resposta.

Na Figura 2 são estudadas possíveis relações que as covariadas possam ter com a concentração total do fósforo.

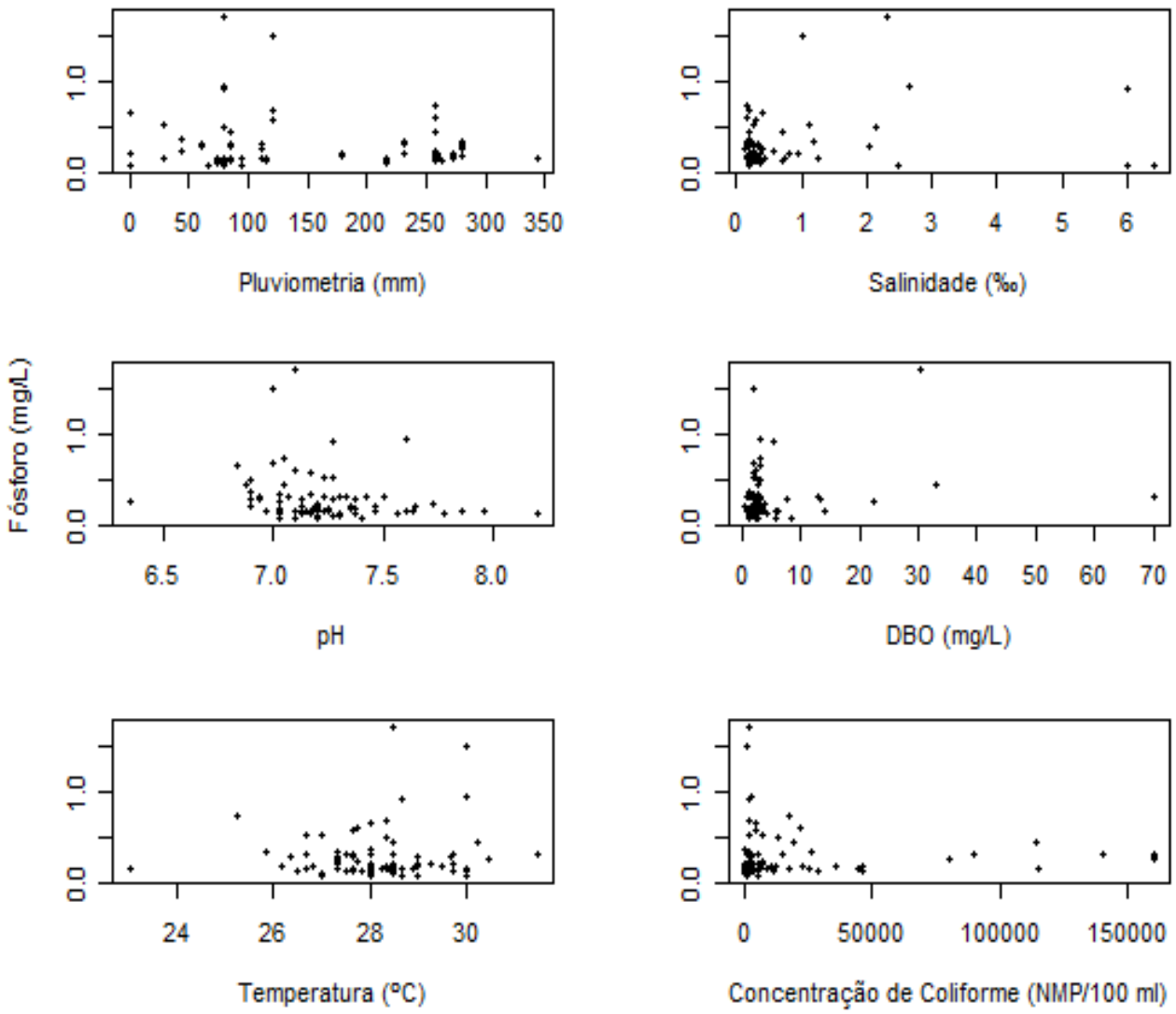

Figura 3 - Relação da variável fósforo com as covariadas em estudo.

Nenhuma relação foi percebida entre a variável de interesse e as covariadas, o que dificulta o ajuste.

Na Tabela 3, a representação matemática de cada fator físico-químico a fim de simplificar na nomenclatura das variáveis consideradas neste relatório.

Tabela 3 - Representação das covariáveis que serão abordadas ao longo do trabalho

Representação dos fatores	Fatores físico-químicos
X_1	Precipitação Total Mensal (Pluviometria)
X_2	Salinidade
X_3	pH
X_4	Demanda bioquímica de oxigênio (DBO)
X_5	Temperatura
X_6	Concentração de coliforme fecais

Obtendo as correlações dos fatores físico-químicos, nota-se que a temperatura e o coliforme possuem as menores correlações com a variável fósforo.

Tabela 4 - Matriz de correlação entre os fatores físico-químicos

	Fósforo	Pluviometria	Salinidade	pH	D.B.O	Temperatura	Coliforme
Fósforo	1	-	-	-	-	-	-
Pluviometria	-0.16	1	-	-	-	-	-
Salinidade	0.23	-0.32	1	-	-	-	-
pH	-0.2	0.02	-0.09	1	-	-	-
DBO	0.21	-0.2	0.09	-0.09	1	-	-
Temperatura	**0.08**	-0.4	0.2	0.06	0.24	1	-
Coliforme	**-0.02**	-0.001	-0.08	-0.28	0.55	0.21	1

Baseado nas correlações, o modelo proposto será composto pelas covariadas pluviometria, salinidade, pH e DBO - que apresentam maior correlação com o fósforo - e suas interações. Uma explicação plausível para a eliminação dessas duas covariadas em uma visão ambiental é a baixa variação da temperatura nas três estações. No caso do coliforme fecais pode-se dizer que não se trata da única fonte de fósforo, uma vez que a atividade agrícola é alta naquele local.

O modelo completo proposto é dado por:

$$\mathbf{y} = \alpha_1 + \alpha_2 + \alpha_3 + \varepsilon, \tag{12}$$

onde

$$\alpha_1 = \beta_0 + \sum_{i=1}^{4} \beta_i \mathbf{x}_i^2 \tag{13}$$

$$\alpha_2 = \sum_{i=1}^{4} \beta_{i+4} \mathbf{x}_i \tag{14}$$

$$\alpha_3 = \sum_{i \neq j} \beta_{ij} \mathbf{x}_i \mathbf{x}_j, i, j = 1,2,3,4. \tag{15}$$

Em que a Equação (13) descreve a parte quadrática do modelo, a Equação (14) descreve a parte linear do modelo e a Equação (15) descreve as possíveis interações do modelo. Será testado o modelo completo, Equação (12), utilizando as três funções de ligações citadas na Tabela 2. Serão inseridas as três funções de ligação em cada modelo final encontrado a fim de perceber qual melhor se adéqua ao modelo.

Os modelos foram comparados pelo coeficiente de Akaike (AIC) (Paula, 2010) e a medida de ajustamento (Turkman, 2000). O AIC desempenha um papel de seleção entre

modelos, quanto menor o valor do AIC melhor o modelo. A medida do ajustamento, conhecida na análise de modelos lineares como coeficiente de determinação, percebe o quanto o modelo está explicando da variabilidade total.

A aplicação do teste t ao modelo completo utilizando a função de ligação inversa num processo iterativo de eliminação concluiu, porém, modelo final apenas com as covariadas significativas dispostas na Tabela 5.

Tabela 5 - Covariadas significantes do modelo final com a escolha da função de ligação inversa

Covariada	Estimativa	Erro Padrão	P-valor
Intercepto	-47,1	13,83	0.0011
$\boldsymbol{X_2}$	26,02	8,44	0.003
$\boldsymbol{X_2^2}$	0,19	0,09	0.041
$\boldsymbol{X_3}$	1,47	0,81	0.07
$\boldsymbol{X_4}$	7,15	1,93	0.0004
$\boldsymbol{X_2X_3}$	-0.2	0.11	0.07
$\boldsymbol{X_2X_4}$	-3.78	1.12	0.001
$\boldsymbol{X_3X_4}$	-0.05	0.02	0.04

Inserindo as funções de ligação logarítmica e identidade no modelo com função de ligação inversa, obtêm-se os coeficientes de Akaike e as medidas de ajustamentos de cada modelo dispostos na Tabela 6, onde se observa que o melhor modelo obtido foi com a função de ligação inversa, o que era esperado.

Tabela 6 – Coeficiente de Akaike e medida de ajustamento com as ligações inversa, logarítmica e identidade

Função de ligação	Inversa	Logarítmica	Identidade
AIC	**-82,42**	-77,22	-73,5
Medida de ajustamento	**0,35**	0,31	0,25

Através do teste t, o modelo final utilizando a função de ligação logarítmica obteve as covariadas significativas dispostas na Tabela 7.

Tabela 7 - Covariadas significantes do modelo final com a escolha da função de ligação logarítmica

Covariada	Estimativa	Erro Padrão	P-valor
Intercepto	7,28	2,53	0.005
$\boldsymbol{X_2}$	-7,87	3,54	0.03
$\boldsymbol{X_2^2}$	-0,13	0,04	0.001
$\boldsymbol{X_3}$	-1,23	0,35	0.007
$\boldsymbol{X_2X_3}$	1,21	0,49	0.01

Neste modelo com ligação logarítmica será inserindo as funções de ligação inversa e identidade, obtêm-se os coeficientes de Akaike e as medidas de ajustamentos de cada modelo na Tabela 8, onde se vê que a função logarítmica é a mais adequada.

Tabela 8 – Coeficiente de Akaike e medida de ajustamento com as ligações inversa, logarítmica e identidade

Função de ligação	Inversa	Logarítmica	Identidade
AIC	-71,95	**-79,32**	-73,49
Medida de ajustamento	0,23	**0,30**	0,25

A partir do teste t, chega-se ao modelo final utilizando a função de ligação identidade, cujo covariadas significativas estão dispostas na Tabela 9.

Tabela 9 - Covariadas significantes do modelo final com a escolha da função de ligação identidade

Covariada	Estimativa	Erro Padrão	P-valor
Intercepto	-47,1	13,83	0.0011
$\boldsymbol{X_2}$	26,02	8,44	0.003
$\boldsymbol{X_2^2}$	0,19	0,09	0.041
$\boldsymbol{X_3}$	1,47	0,81	0.07

Inserindo as funções de ligação inversa e logarítmica no modelo final encontrado com ligação identidade, obtêm-se os coeficientes de Akaike e as medidas de ajustamentos de cada modelo dispostos na Tabela 10, onde se vê que a função de ligação logarítmica teve o melhor comportamento.

Tabela 10 – Coeficiente de Akaike e medida de ajustamento com as ligações inversa, logarítmica e identidade

Função de ligação	Inversa	Logarítmica	Identidade
AIC	-71,25	**-74,09**	-73,49
Medida de ajustamento	0,23	**0,25**	0,25

Com os resultados da tabela 6, 8 e 10 conclui-se que a função de ligação canônica da Gama foi a que melhor se adaptou ao modelo, constando o melhor AIC e medida de ajustamento, sendo assim, será trabalhado nesse texto o modelo com função de ligação inversa.

Logo o modelo final que descreve a concentração do fósforo total com os coeficientes estimados fornecida na Tabela 5 é dado por:

$$ft = \left(-47{,}1 + 26{,}02\mathbf{X_2} + 0{,}19\mathbf{X_2^2} + 1{,}47\mathbf{X_3} + 7{,}15\mathbf{X_4} - 0{,}2\mathbf{X}_2\mathbf{X}_3 - 3{,}8\mathbf{X}_2\mathbf{X}_4 - 0{,}05\mathbf{X}_3\mathbf{X}_4\right)^{-1}, \qquad (12)$$

em que ft é o valor ajustado do modelo.

Será efetuada uma análise de diagnóstico a fim de validar e explicar se de fato o modelo final encontrado está propício para mostrar bons resultados e interpretações referentes à concentração de fósforo no estuário do rio Goiana e região costeira adjacente.

Na Figura 4 percebe-se graficamente certo crescimento nos resíduos studentizados na medida em que os valores ajustados aumentam, mas pouco significativo para de fato afirmar que existe uma heterocedasticidade, comprovado pelo teste de Breusch-Pagan (Yan, 2009) cujo p-valor resultou em 0,37. Assim, não rejeitando a hipótese da homocedasticidade nos resíduos. Também se pode notar que na Figura 4 os pontos 42 e 65 são candidatos para serem *outliers*. Na Figura 5 mostra que o modelo está bem adequado para a distribuição Gama na parte aleatória, uma vez que os resíduos estão se comportando de forma aleatória no sentido da reta e por estar dentro do envelope. Na Figura 6 observa-se que a função de ligação escolhida está adequada para o modelo proposto, notando certo comportamento crescente ao longo da reta na relação de uma variável aditiva com o preditor linear.

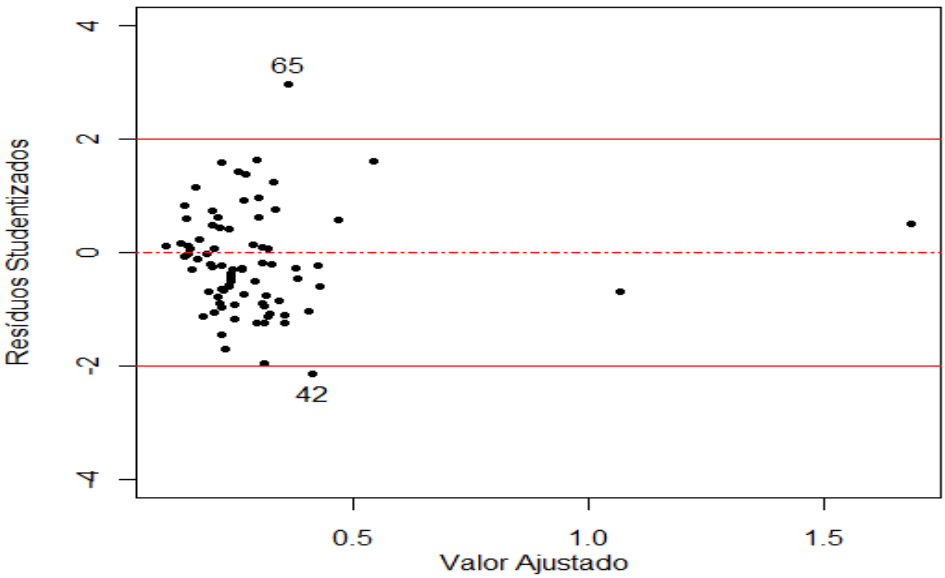

Figura 4 – Gráfico da relação entre o resíduo studentizado com os valores ajustados para percepção de homocedasticidade dos resíduos e identificação de pontos aberrantes no modelo final.

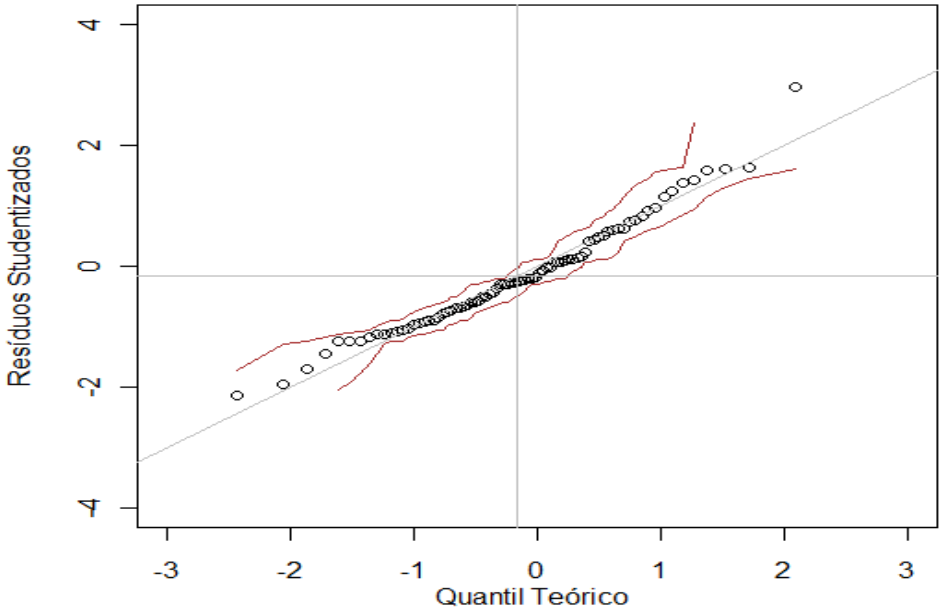

Figura 5 – Gráfico Q-Q plot para analisar a escolha da distribuição da variável resposta

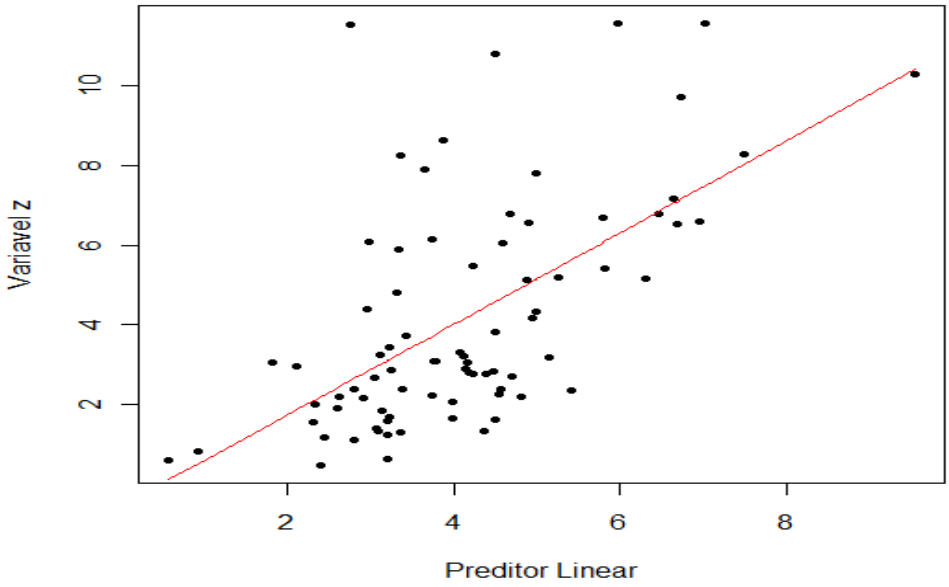

Figura 6 – Gráfico da relação de uma variável aditivada com o preditor linear a fim de perceber a adequacidade do ajuste para a função de ligação inserida

Como na Figura 4 foi identificado dois pontos *outliers*, são eles o 42 e 65. Na Tabela 11 foi verificado o aumento no coeficiente de Akaike e na medida de ajustamento a partir do

momento que é retirado individualmente o ponto 42, sem perder nenhuma inferência. Quando retirado o ponto 65 o modelo proporcionou uma maior explicação da variabilidade total, mas perdendo três covariadas. Uma explicação para o ponto 65 é que a fonte de fósforo não vem do esgoto nesse canal e sim das terras que contém fósforo originado de atividades agrícolas arrastadas para o estuário através do grande volume de chuva. Retirando ambos os pontos, a variabilidade total e o coeficiente de Akaike não mostrou grande diferença quando retirado apenas o ponto 65.

Tabela 11 – Mudanças no Coeficiente de Akaike e medida de ajustamento quando os pontos *outliers* são removidos do modelo

Observação retirada	AIC	Medida de ajustamento
42	-84,06	0,38
65	-101,65	0,38
42 e 65	-100,31	0,39

6. CONCLUSÃO

Foram testadas três diferentes funções de ligação a fim de compará-las em termos de qualidade do ajustamento, a função de ligação que foi selecionada através do coeficiente de Akaike foi a função inversa, as outras funções obtiveram resultados não tão distante se comparada com a inversa.

Concluiu-se do modelo final, Equação (12), que se a salinidade crescer conjuntamente com o DBO, mantido fixo o pH, a concentração de fósforo total é minimizada, uma vez que as observações relacionadas ao pH indicaram variações pouco consideráveis neste local. A salinidade foi a única covariada que ao ajustar a polinomial quadrática manteve o grau dois, de forma significativa no modelo. Pode-se concluir, também, que diretamente responsáveis são as interações dos fatores físico-químicos - salinidade, DBO e pH – no aumento desta concentração.

Com base na satisfatória modelagem para a concentração de fósforo total presente nos estuários do rio Goiano e região costeira adjacente, e a forte contribuição das interações dos fatores abióticos aqui verificada, fazem-se interessantes futuras pesquisas referentes à produtividade, uma vez que o fósforo está diretamente relacionada a alimentação de algumas espécies da região estuarina

7. REFERÊNCIAS BIBLIOGRÁFICAS

BARLETTA, Mario; Costa, M., F.. *Living and non-living resources exploitation in a tropical semi-arid estuary*. Journal of Coastal Research, 2009.

DEMÉTRIO, C. G. Borges. *Modelos Lineares Generalizados em Experimentação Agronômica*. ESALQ/USP, 2002.

DRING, M. J..*The biology of marine plants*. Cambridge University, 1992

ELLIOT, M; DS McLUSKY. 2002. The need definitions in understanding estuaries. Estuar. Coast. Shelf Sci. 55: 815-827.

FIDEM. *Região Metropolitana do Recife: Reservas ecológicas*. Recife, Governo do Estado de Pernambuco. Secretearia de Planejamento do Estado de Pernambuco, Fundação de Desenvolvimento da Região Metropolitana do Recife. I+108p. 1987.

MIRANDA, B. M., CASTRO, B. M., KJERFEVE, B. Princípios de oceanografia física de estuários. São Paulo: EDUSP, 2002. 441p.

MONTGOMERY, Douglas C. *et al*. *Introduction to Linear Regression Analysis.* 4 ed. New Jersey: Wiley, 2006.

PAULA, Gilberto A. *Modelo de Regressão com apoio computacional*. IME/USP, 2010.

ROSS, S. M. *Simulation*. Elsevier, 2006.

SANTOS, Maria de Lourdes Souza et al. *Estudo das diferentes formas de fósforo nas águas da plataforma continental do Amazonas*. Quim. Nova, v.30, n.3, p.569-573, 2007.

TURKMAN, M. A. A.; SILVA, G.L.. *Modelos Lineares Generalizados*: da teoria à prática.VIII Congresso Anual da Sociedade Portuguesa de Estatística. Lisboa, 2000.

WEISBERG, Sanford. *Applied Linear Regression.* 3 ed. New Jersey: Wiley, 2005.

YAN, Xin; SU, Xiao Gang. *Linear Regression Analysis*: Theory and computing. Cingapure: World Scientific, 2009.