Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística

Uma proposta para análise de dados com correlação espacial e temporal

Flávia Maria de Toledo Pedroso

Orientadora: Maria Cecília Mendes Barreto
Co-Orientadora: Maria Sílvia de Assis Moura

Dissertação apresentada ao Departamento de Pós-Graduação em Estatística da Universidade Federal de São Carlos - UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos
Novembro/2007

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

P372pa

Pedroso, Flávia Maria de Toledo.
Uma proposta para análise de dados com correlação espacial e temporal / Flávia Maria de Toledo Pedroso. -- São Carlos : UFSCar, 2007.
108 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Correlação. 2. Modelos lineares generalizados. 3. Equações de estimação generalizadas. 4. Geoestatística. 5. Stepwise. 6. R (Linguagem de programação de computador). I. Título.

CDD: 519.537(20ª)

Dedicatória

Aos meus pais,
Moacyr Pedroso (*in memoriam*) e
Vera Lúcia de Toledo Pedroso,

aos meus irmãos,
Gustavo José de Toledo Pedroso
Leandro José de Toledo Pedroso

aos meus avós,
Cassiano Pereira P. de Toledo (*in memoriam*) e
Hermínia C. de Toledo (*in memoriam*),
Maria de Oliveira Pedroso (*in memoriam*)

à
Nina

Fontes de verdadeiro amor, carinho, incentivo, dedicação e exemplos de vida.
Estímulos em minha vida.

Agradecimentos

Às queridas professoras Maria Cecília Mendes Barreto e Maria Sílvia de Assis Moura pela orientação, amizade, confiança e incentivo que tão generosamente me ofereceram ao longo dos estudos.

Ao professor Francisco Chiaravalloti Neto que não apenas forneceu os dados utilizados neste trabalho, como ainda importantes informações e esclarecimentos sobre as características do *Aedes Aegypti*.

Aos professores Lael Almeida de Oliveira e Paulo Milton Barbosa Landim pelas sugestões e pelo apoio.

Aos professores do DEs, em especial Vera Lúcia Damasceno Tomazzela, Jorge Achar e José Galvão Leite, pelos ensinamentos que deles recebi.

Aos queridos amigos do mestrado: Adriano Kamimura Suzuki, Daniela Parreira, Camila Lima, Glaucy Parolin, Diomedes Pael, Olympio Teixeira Neto, Luis Ernesto Bueno e Sabrina Caetano, pela convivência, pela troca de conhecimentos, e pelo grande apoio e incentivo nestes anos.

À querida funcionária do DEs Dona Luiza Maria da Silva pelo carinho, apoio, amizade e dedicação a todos os alunos e professores do departamento, um exemplo de vida.

Às queridas tias Cleide de Toledo Gomes, Maria Aparecida Pedroso Miyahara, Maria Cecília Pedroso, Maria José Pedroso Mayr, Amethista Pedroso, Ana Grabriela Pedroso, Vera Pedroso Hull, Vera Brito Pedroso, Dora M. R. Pedroso e aos queridos tios Alencar Pedroso e José Joaquim Pedroso pelo grande incentivo e apoio.

Aos primos, em especial Sandra, Cassiano, Aninha e Étore, Márcio e Virgínia, Giulia, Rodrigo, Priscila, Fabiana, Gabrielinha, Mariozito, Tiago, Rick, Gordon e Bárbara pela grande amizade, carinho, apoio e por se fazerem presentes nos momentos difíceis.

À Vergínia Hilário de Oliveira e Rosana Aparecida de Oliveira Madeira pela amizade, carinho e dedicação ao longo da vida.

Ao Paulo Henrique de Avelino Rodrigues pela paciência nas horas em que estive distante, pelo apoio, carinho e dedicação.

À educadora e grande amiga Vânia Cerqueira Leite pelo grande incentivo e carinho.

Às amigas Iracy Chiodi Galvanini, Daniela Galvanini, Gladis F. A. Panigalli, Sônia A. Borges Ignácio, Meire R. Barnese, Fabiana Romano Bressan, Salete Lhamas, Renata Ventura e Iara Borges Leal pela amizade, carinho, apoio e por se fazerem presentes tanto nos momentos felizes como nos momentos difíceis.

Aos amigos da Fundação Sidelma D. Leite de Souza, Daniel Fernando Christianini e Liliam Nanci Carlos pela amizade, convivência e apoio.

Ao querido amigo Dimas Fernando B. de Oliveira pelo grande apoio e incentivo.

Ao querido professor Dr. Hilton Aparecido Garcia pela carta de recomendação, pela confiança e o grande apoio.

À Dra. Cleusa Camilo Atique e ao Dr. Abdala Atique pela amizade e o carinho dedicados a mim e minha família.

À professora Maria Hermínia Marques Leite pela amizade, incentivo e confiança dispensados tanto no início da minha carreira quanto no decorrer da minha vida profissional.

Aos grandes amigos e professores da FATEC Rafael Garcia M. Filho (*in memoriam*), Sérgio Lukine e Aramis M. C. de Mendonça.

Aos queridos professores que me proporcionaram crescimento pessoal e profissional Ana Maria Cardoso Ventura, Ademir José Ventura, Maria Isabel Dário Oliboni, Dolores Pintor de Arruda, Maria Eliza Milani Sarkis, Maria Aparecida Morijo, Sérgio Luiz Bertoncello, Ismael Antonio Silva, Suzana Abrunhosa, Yvette Eliza P. Grizzo, Maria de Lourdes M. de Almeida Prado, Terezinha B. Buffo e “Tia” Arlete Ortigoza.

À querida professora Ana Helena Neuber de Oliveira pelo grande apoio.

Ao professor Evandro Antonio Bertoluci pelo incentivo, apoio e confiança.

Resumo

É muito comum, em diversas áreas, o estudo da ocorrência de um fenômeno ao longo do tempo. Neste caso, se utilizarmos a teoria de Modelos Lineares Generalizados para analisarmos o objeto de interesse, teremos como conseqüência inferências incorretas dos parâmetros regressores e estimadores ineficientes, uma vez que a principal característica desta teoria é considerar as variáveis aleatórias como sendo respostas independentes.

Quando a variável resposta é observada ao longo do tempo, pode haver uma correlação entre as observações e isso deve ser levado em consideração na estimação dos parâmetros. Para incorporarmos esta dependência temporal podemos utilizar a teoria das Equações de Estimação Generalizadas, proposta por Liang & Zeger, 1986, como uma extensão dos Modelos Lineares Generalizados para computar a correlação existente entre as observações.

Além da correlação temporal, pode haver, ainda, uma correlação espacial e, neste caso, podemos utilizar a teoria da Geoestatística para estimarmos o alcance de correlação das amostras ao longo de uma região de estudo, bem como para identificarmos se há uma direção privilegiada de variabilidade do fenômeno analisado, dados importantes não revelados quando utilizamos as teorias da estatística clássica.

Nesta dissertação aplicamos as metodologias acima citadas para tentar explicar a presença e o comportamento de fêmeas *Aedes (Stegomyia) aegypti* capturadas por armadilhas adulticidas na cidade de Mirassol/SP, com o objetivo de colaborar na busca de métodos mais precisos para contenção da disseminação da dengue.

Abstract

In several research areas, the study of the occurrence of a phenomenon over a period of time is very common. In this case, if we use the theory of Generalized Linear Models to analyze the subject of interest, we'll have, as a consequence, incorrect inferences concerning regression parameters and inefficient estimators, since considering the random variables as independent responses, is the main characteristic of this theory.

When the variable response is observed over time, there can be a correlation between the observations and that must be taken into consideration on the estimation of the parameters. To incorporate this temporal dependence, we can use the theory of Generalized Estimating Equations, proposed by Liang & Zeger, 1986, as an extension of the Generalized Linear Models, to compute the correlation between the observations.

Besides the temporal correlation, there can even be a space correlation and, in this case, we can use the theory of Geostatistic to estimate the reach of the correlation of samples over a study region, as well as to identify whether there is a privileged direction of variability of the studied phenomenon, important data not revealed when we utilize the theories of classic statistic.

In this dissertation, we applied the methodologies mentioned above to try to explain the presence and the behavior of the female *Aedes (Stegomyia) aegypti* captured by adulticed traps in the city of Mirassol/SP, with the goal of helping on the search of more precise methods to contain the dissemination of dengue.

Sumário

1 - INTRODUÇÃO

A década de 70 evidenciou um grande avanço computacional que favoreceu a utilização de processos iterativos na estimação dos parâmetros em modelos que apresentavam essa exigência, tornando possível, assim, o surgimento na literatura de modelos de regressão mais elaborados.

Segundo DEMÉTRIO (2002), em 1972, Nelder & Wedderburn propuseram a teoria de Modelos Lineares Generalizados - MLG, na qual, dentre outras características, destacamos que a variável resposta tem uma distribuição pertencente à família exponencial e a amostra é composta por observações independentes.

Entretanto, esta suposição de independência pode não ser verificada em determinadas situações, como no caso de pesquisas que envolvem o estudo de dados correlacionados, provenientes de medidas repetidas (dados longitudinais) ou de agrupamentos.

Esta correlação deve ser incorporada por métodos apropriados de análise de dados. Se a pesquisa em estudo possui dados correlacionados e se eles forem tratados como independentes, teremos como conseqüência: inferências incorretas dos parâmetros da regressão devido a erros padrões subestimados; estimadores ineficientes, ou seja, haverá um maior erro quadrático médio nos estimadores dos parâmetros da regressão.

Neste sentido, conforme explicitado por JOHNSTON et al. (1996), foi proposto o uso das Equações de Estimação Generalizadas - EEG - (Liang & Zeger, 1986), um método de estimação dos parâmetros do modelo da regressão para acomodar a suposição de dados correlacionados.

Em muitas situações, além de uma dependência temporal, a variável de interesse pode apresentar uma interdependência espacial. Com o objetivo de analisar a variação da variável resposta em relação às coordenadas de localização da amostra e, ainda, predizer o valor da variável dependente em locais não amostrados, Matheron em 1963, como apresentado em LANDIM (2003), formulou a teoria de Geoestatística que, através de uma função chamada semivariograma, nos permite conhecer o comportamento desta variação no espaço a partir da estimativa do alcance de dependência entre amostras vizinhas e da identificação da direção de maior espalhamento do fenômeno em estudo, podendo gerar um modelo com maior poder preditivo.

Neste estudo apresentaremos uma análise estatística e seus fundamentos teóricos de um conjunto de dados provenientes do projeto de pesquisa "*Estudo da relação entre indicadores entomológicos para Aedes (Stegomyia) aegypti obtidos de armadilhas adulticidas, de oviposição e de coleta de adultos, em área da região noroeste do estado de São Paulo*"

(CHIARAVALLOTI-NETO et al. (2004), Faculdade de Medicina de São José do Rio Preto), cujo objetivo geral foi avaliar as relações entre os indicadores entomológicos obtidos através de armadilhas adulticidas, de oviposição e de coletas de adultos e os indicadores climáticos em uma área da região noroeste do estado de São Paulo. Mais especificamente, nesta dissertação trabalhamos com os dados da armadilha adulticida *Mosquitrap*, desenvolvida por EIRAS (2002), cuja utilização foi proposta com o objetivo de se estudar o estado fisiológico do mosquito *Aedes*, bem como a paridade das fêmeas, dados fundamentais para caracterização biológica dos vetores da dengue, já que propiciam conhecimentos sobre sua capacidade de infectar-se e transmitir o vírus (BARATA et al. (2001)). Além disso, identificando uma determinada quantidade de insetos capturados nestas armadilhas no intra e peridomicílio das residências amostradas, podemos prever uma possível epidemia de dengue e, ainda, relacionar a presença do *Aedes* com variáveis meteorológicas.

Nosso objetivo é, então, pesquisar modelos que exprimam a relação existente entre o número de *Aedes* capturadas com as variáveis meteorológicas e modelos que expliquem o comportamento espacial deste fenômeno com a finalidade de colaborar na busca de métodos mais precisos para contenção da disseminação da dengue.

Inicialmente, para explicar a relação existente entre o número de fêmeas *Aedes* capturadas pelas armadilhas adulticidas e os indicadores climáticos, usamos Modelos Lineares Generalizados (MLG) e, posteriormente, por este conjunto de dados possuir uma dependência temporal, apresentamos como proposta de modelagem o uso da teoria de Equações de Estimação Generalizadas (EEG), ambas teorias da estatística clássica. Além da dependência temporal, há evidências da existência de uma correlação espacial e, neste sentido, utilizamos a teoria da Geoestatística para estudar a variação deste fenômeno no espaço, ou melhor, estudar sua estrutura de dependência espacial para que atitudes de controle mais eficientes possam ser efetuadas.

Nos modelos baseados em MLG e EEG, para identificarmos as possíveis variáveis meteorológicas que exercem influência na captura dos mosquitos pela armadilha, utilizamos a metodologia de seleção de covariáveis stepwise via critério valor *p* ou nível descritivo do teste. Em MLG, ao considerar a quadra como bloco, não encontramos na literatura um programa/função disponível quando um conjunto de variáveis preditoras fazem parte do modelo minimal no procedimento de seleção de variáveis. Com relação ao ajuste de modelos via EEG, não encontramos programas em que a metodologia de seleção de variáveis já estivesse implementada. Neste contexto, desenvolvemos funções específicas em *R* para automatizar o procedimento de seleção de variáveis stepwise via critério valor *p* a partir de um modelo

minimal, com covariáveis fixas ou não, tanto para MLG como EEG.

Desta forma, no Capítulo 2 apresentamos uma revisão da literatura dos Modelos Lineares Generalizados, Equações de Estimação Generalizadas e Geoestatística.

No Capítulo 3 descrevemos o material e métodos utilizados, explicitando os parâmetros das funções stepwise desenvolvidas em *R*.

O Capítulo 4 relata os resultados e principais conclusões e, no Capítulo 5, apresentamos uma proposta da continuação deste trabalho.

2 - REVISÃO DE LITERATURA

2.1 Modelos de Regressão

Um modelo de regressão é uma equação matemática do tipo $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ que descreve um fenômeno de interesse, avaliado em termos de uma variável, ou vetor, aleatória **Y**, de dimensão *n* x 1, em função *p* de variáveis não estocásticas conhecidas (covariáveis), **X**, de dimensão *n* x *p*, e de um efeito aleatório, ε, de dimensão *n* x 1, não identificado, incorporado ao modelo para expressar erros de medidas, efeitos de variáveis não incluídas ou, ainda, variabilidade natural inerente ao fenômeno estudado. Vale salientar que esta equação é linear nos parâmetros β desconhecidos.

O principal objetivo dos modelos de regressão é estimar o efeito das covariáveis (ou variáveis independentes ou preditoras ou, ainda, explicativas) $\mathbf{X} = (X_1, X_2, ..., X_p)$ sobre uma variável resposta (ou variável dependente) **Y** de interesse experimental, discreta ou contínua, ou melhor, descrever o relacionamento de **Y** com **X**, para, assim, prevermos valores futuros de **Y** ou avaliarmos o efeito destas variáveis explicativas sobre a resposta ou, ainda, descrevermos a estrutura dos dados. Desta forma, utilizando modelos de regressão mais gerais, temos que o valor esperado de **Y** é uma função das covariáveis $\mathbf{X} = (X_1, X_2, ..., X_p)$, ou seja, $E[\mathbf{Y}/\mathbf{X}] = f(\mathbf{X})$. Este relacionamento pode ser expresso por uma equação linear ou uma função não linear.

Em modelos de regressão, se considerarmos as respostas independentes e com distribuição pertencente à família exponencial, uma das alternativas é utilizarmos a teoria dos **Modelos Lineares Generalizados**. Se há dependência entre as observações, dados correlacionados provenientes de medidas repetidas (dados longitudinais) ou de agrupamentos, usamos as **Equações de Estimação Generalizadas**. Quando levamos em consideração a interdependência espacial entre eventos, devemos utilizar a teoria de **Modelos de Regressão Espacial,** como por exemplo a **Geoestatística** que incorpora na modelagem a dependência espacial dos dados.

2.1.1 Modelos Lineares Generalizados

Durante muito tempo, os modelos normais lineares foram utilizados para descrever a maioria dos fenômenos aleatórios. Mesmo quando o fenômeno sob estudo não apresentava uma resposta para a qual fosse razoável a suposição de normalidade, tentava-se algum tipo de transformação no sentido de alcançar a normalidade procurada.

Com o avanço computacional ocorrido a partir de 1970, que permitiu a implementação de processos iterativos para a estimação dos parâmetros do modelo, diversas teorias foram propostas na literatura. Em 1972, Nelder e Wedderburn introduziram a teoria de

Modelos Lineares Generalizados (MLG) que podem ser utilizados para analisar dados discretos ou contínuos, uma vez que a variável resposta tenha uma distribuição pertencente à família exponencial. Outra característica importante a ser citada é que, assim como em modelos lineares clássicos, as respostas aqui observadas também são consideradas independentes.

Diversos autores, entre eles DEMÉTRIO (2002) e PAULA (2004), apresentam os principais conceitos de Modelos Lineares Generalizados, que agora vamos introduzir.

Considere a variável aleatória Y_i, $i = 1,...,n$, como sendo respostas independentes. Os MLG são caracterizados por:

a) **componente aleatório**: representado por um conjunto de variáveis aleatórias Y_i, $i = 1,...,n$, independentes provenientes de uma mesma distribuição de probabilidade pertencente à família exponencial na forma canônica com médias μ_1, μ_2, μ_3,...., μ_n, ou seja, $E(Y_i)= \mu_i$, $i = 1,..., n$, um parâmetro constante de escala, conhecido, $\phi > 0$, e que depende de um único parâmetro θ_i, chamado canônico ou natural. A função densidade probabilidade (f.d.p.) de Y_i é dada por:

$$f(y_i ; \theta_i , \phi) = exp\left\{ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i ; \theta) \right\} I_A(y) \qquad \textbf{(1)}$$

sendo b(.) e c(.) funções conhecidas. Em geral, $a_i(\phi) = \phi / w_i$, sendo w_i pesos a priori.

A família exponencial é composta pelas distribuições: Binomial, Poisson, Binomial Negativa, Gama, Normal, Normal Inversa ou Inversa Gaussiana, entre outras.

Temos, ainda, que:

$$E(Y_i) = \mu_i = b'(\theta_i)$$

$$Var(Y_i) = a_i(\phi) b''(\theta_i) = a_i(\phi) V(\mu_i) = a_i(\phi) V_i,$$

onde:

$V_i = d\mu_i / d\theta_i$ é chamada função de variância. Como depende unicamente da média, temos que o parâmetro natural pode ser expresso como:

$\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i)$, onde $q(\mu_i)$ é uma função conhecida de μ_i.

b) **componente sistemático**: envolve as variáveis explicativas que entram no modelo na forma de uma soma linear de seus efeitos:

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j = \mathbf{x}_i' \boldsymbol{\beta} \quad ou \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \text{ onde:}$$

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n)'$ é a matriz do modelo;

$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3,..., \beta_p)'$ é o vetor de parâmetros desconhecidos;

$\eta = (\eta_1, \eta_2,...,\eta_n)'$ é o preditor linear.

c) **<u>função de ligação</u>**: é uma função que liga o componente aleatório ao componente sistemático, ou seja, relaciona a média da variável resposta ao preditor linear, isto é,

$$\eta_i = g(E(Y_i)) = g(\mu_i)$$

A inversa da função de ligação é chamada de **função média**: $g^{-1}(\eta_i) = \mu_i$.

Se a função de ligação é escolhida de tal forma que $g(\mu_i) = \theta_i$, o preditor linear modela diretamente o parâmetro canônico e tal função de ligação é chamada **ligação canônica**.

Para um conjunto de observações independentes $y_1, y_2, ..., y_n$, o logaritmo da função de verossimilhança $\ell(\boldsymbol{\theta}, \phi; \mathbf{y})$ é dado pela soma das contribuições individuais:

$$\ell(\boldsymbol{\theta}, \phi; \mathbf{y}) = \ln L(\boldsymbol{\theta}, \phi; \mathbf{y}) = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i; \phi) = \ell(\theta_i, \phi; y_i), \qquad \textbf{(2)}$$

onde $L(\boldsymbol{\theta}, \phi; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \exp\left\{\sum_{i=1}^{n} \frac{1}{a_i(\phi)}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i; \phi)\right\}$ é a função de verossimilhança; e $f(y_i; \theta_i, \phi) = exp\left\{\frac{1}{a_i(\phi)}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i; \theta)\right\} I_A(y)$ é a densidade da família exponencial.

Seja, então, o logaritmo da função de verossimilhança definido por:

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} \ell(\mu_i; y_i),$$

onde $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \mathbf{x}_i'\boldsymbol{ß}$.

Para o modelo saturado, onde o número de observações, *n*, é igual ao número de parâmetros, *p*, a função $\ell(\boldsymbol{\mu}; \mathbf{y})$ corresponde a:

$$\ell(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} \ell(y_i; y_i)$$

e a estimativa de máxima verossimilhança de μ_i é dada por $\hat{\mu}_i^0 = y_i$, $i = 1, ..., n$.

Quando temos $p < n$, denotaremos o logaritmo da função de verossimilhança de $\ell(\boldsymbol{\mu}; \mathbf{y})$ por $\ell(\hat{\boldsymbol{\mu}}, \mathbf{y})$ e a estimativa de máxima verossimilhança de μ_i é dada por $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$, onde $\hat{\eta}_i = \mathbf{x}_i'\hat{\boldsymbol{ß}}$.

A estimação dos parâmetros de um MLG é feita pelo método de máxima verossimilhança. Assim, derivamos o logaritmo da função de verossimilhança em relação ao β,

$\frac{\partial \ell(\theta_i; y_i, \phi)}{\partial \beta_j}$, j = 1, 2, ..., p e obtemos um sistema de funções não lineares $U(\beta_j)$, j = 1, 2,..., p, denominada *função escore*, expressa por:

$$U(\beta_j) = \sum_{i=1}^{n} \frac{\partial \ell(\theta_i; y_i, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{\phi} x_{ij} W_i \frac{d\eta_i}{d\mu_i} [y_i - \mu_i] \text{ (Apêndice A – A1)},$$

na qual $W_i = \frac{w_i}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2$.

Podemos escrever a *função escore* na forma vetorial:

$$U(\mathbf{ß}) = \frac{\partial \ell(\theta; \mathbf{y}, \phi)}{\partial \mathbf{ß}} = \frac{1}{\phi} \mathbf{X'W\Delta(y - \mu)},$$

onde **X** é uma matriz n x p de posto completo, cujas linhas serão denotadas por $\mathbf{x}'_i$, i = 1, ..., n, **W** = diag{W_1, ..., W_n} é a matriz de pesos, Δ = diag{$d\eta_1/d\mu_1$, ..., $d\eta_n/d\mu_n$} = diag{$g'(\mu_1)$, ..., $g'(\mu_n)$}, **y** = {y_1, ..., y_n}' e $\boldsymbol{\mu}$ = {μ_1, ..., μ_n}'.

Como vimos, as funções $U(\beta_j)$ não são lineares e, quando igualadas a zero, devem ser resolvidas por processos iterativos, como por exemplo, o método escore de Fisher que utiliza a matriz de informação esperada de Fisher e a função escore para estimar os β's.

Para obtermos a matriz de informação de Fisher, precisamos calcular a esperança da segunda derivada do logaritmo da função de verossimilhança:

$$\mathfrak{I}(\beta)_{jk} = E\left(\frac{\partial^2 \ell(\theta; y, \phi)}{\partial \beta_j \partial \beta_k} \right) = E(U_j U_k) = \sum_{i=1}^{n} \frac{1}{\phi} x_{ij} W_i x_{ik} \text{ (Apêndice A – A2)},$$

onde $W_i = \frac{w_i}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2$.

Desta forma, podemos escrever a informação de Fisher na forma matricial:

$$\mathfrak{I}(\mathbf{ß}) = \frac{1}{\phi} \mathrm{X'WX},$$

onde **X** é a matriz do modelo, **W** = diag{W_1, ..., W_n} é a matriz de pesos e ϕ é o parâmetro de dispersão.

A estimativa de máxima verossimilhança de β, $\hat{\mathbf{ß}}$, é obtida através do processo iterativo de Newton-Raphson, expandindo-se a função escore **U(β)** em torno de um valor inicial $\beta^{(0)}$, tal que

$$\mathbf{U(\beta) \cong U(\beta^{(0)}) + U'(\beta^{(0)})(\beta - \beta^{(0)})},$$

onde **U'(β)** representa a primeira derivada **U(β)** com respeito a β.

Repetindo-se o procedimento acima, chegamos ao processo iterativo

$$\beta^{(m+1)} = \beta^{(m)} + \{-\mathbf{U'}(\beta^{(m)})\}\mathbf{U}\,(\beta^{(m)}),\ m = 0, 1, ...$$

Já que a matriz $-\mathbf{U'}(\beta)$ pode não ser positiva definida, aplicaremos método *scoring* de Fisher, substituindo esta matriz pelo seu correspondente valor esperado, resultando em:

$$\beta^{(m+1)} = \beta^{(m)} + \mathfrak{I}^{-1}(\beta^{(m)})\mathbf{U}\,(\beta^{(m)}),\ m = 0, 1, ...$$

Fazendo-se algumas alterações, chegaremos a um processo iterativo de mínimos quadrados reponderados:

$$\beta^{(m+1)} = (\mathbf{X'W}^{(m)}\mathbf{X})^{-1}\mathbf{X'}\,\mathbf{W}^{(m)}\mathbf{z}^{(m)},\ m = 0, 1, ..., \qquad \mathbf{(3)}$$

na qual $\mathbf{z}^{(m)} = \eta^{(m)} + \Delta^{(m)}\,(\mathbf{y} - \mu)^{(m)}$ é chamada variável dependente ajustada.

Iniciamos este processo iterativo especificando uma estimativa inicial para $\beta^{(0)}$ e procedemos alterando-a sucessivamente até obtermos a convergência e, assim, $\hat{\boldsymbol{ß}} = \boldsymbol{ß}^{(m+1)}$.

Com estes resultados, podemos resumir os passos do algoritmo de estimação como segue:

(1) obter as estimativas $\eta_i^{(m)} = \sum_{j=1}^{p} x_{ij}\beta_j^{(m)}$ e $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$;

(2) obter a variável dependente ajustada $z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)})$ e os pesos $W_i^{(m)} = \dfrac{w_i}{V(\mu_i^{(m)})\left[g'(\mu_i^{m})\right]^2}$;

(3) calcular $\beta^{(m+1)} = (\mathbf{X'W}^{(m)}\mathbf{X})^{-1}\mathbf{X'}\,\mathbf{W}^{(m)}\mathbf{z}^{(m)}$, voltar ao passo (1), fazer $\beta^{(m)} = \beta^{(m+1)}$ e repetir o processo até a convergência.

Temos, sob condições gerais de regularidade, que $\hat{\boldsymbol{ß}}$ é um estimador consistente e eficiente de β e que, conforme $n \to \infty$, $\sqrt{n}\left(\hat{\boldsymbol{ß}} - \boldsymbol{ß}\right) \xrightarrow{d} N_p\left(0, \phi^{-1}\mathbf{S}^{-1}(\boldsymbol{ß})\right)$, onde $\Sigma(\beta)$ é uma matriz positiva definida, $\Sigma(\boldsymbol{ß}) = \lim_{n\to\infty} \dfrac{\mathfrak{I}(\boldsymbol{ß})}{n}$ e $\mathfrak{I}(\beta)$ não contém aqui o multiplicador ϕ.

Para obtermos os erros padrões, intervalos de confiança e testes de hipóteses para os $\hat{\boldsymbol{ß}}$'s, precisamos estimar do parâmetro de dispersão ϕ.

Vale salientar que os parâmetros $\hat{\phi}$ e $\hat{\boldsymbol{ß}}$ são ortogonais, já que $E[\partial^2 \ell\,(\beta,\phi;\mathbf{y})/\partial\beta_j\,\partial\phi] = 0$. Este fato garante a independência assintótica entre $\hat{\phi}$ e $\hat{\boldsymbol{ß}}$.

Os métodos mais utilizados para estimação de ϕ são: método da máxima verossimilhança, método dos momentos e uma estimativa baseada na estatística de Pearson χ^2.

(a) método da máxima verossimilhança: derivando-se o logaritmo da função de verossimilhança apenas com relação ao parâmetro ϕ e igualando a zero, obtemos a estimativa de máxima verossimilhança para ϕ: $\partial\, \ell(\phi, \boldsymbol{\beta}; \mathbf{y}) / \partial\phi = 0$;

(b) método dos momentos: nos fornece uma outra estimativa não consistente para ϕ, através de:

$$\tilde{\phi} = \frac{D_p}{n-p},$$

onde D_p é a deviance do modelo sob estudo, n é o número de observações e p é o número de parâmetros do modelo sob estudo.

Este método baseia-se no fato de que $S_p \sim \chi^2_{n-p}$, o que nem sempre é verdade.

Uma estimativa considerada melhor que a anterior é

$$\tilde{\tilde{\phi}} = \frac{D_m}{n-m},$$

onde D_m é a deviance do modelo maximal, n é o número de observações e m é o número de parâmetros do modelo maximal. Neste caso, espera-se que a *scaled deviance* S_m tenha um valor mais próximo da esperança da qui-quadrado, ou seja,

$$E(S_m) = \frac{1}{\tilde{\tilde{\phi}}} E(D_m) \cong n-m.$$

(c) estimativa baseada na estatística de Pearson χ^2 generalizada: é dada por:

$$\phi^* = \frac{\chi^2}{n-m},$$

onde n é o número de observações e m é o número de parâmetros do modelo maximal.

Vale salientar que esta estatística nem sempre é não viesada, mas, é consistente.

A qualidade do ajuste de um MLG é avaliada através da função desvio (deviance), dada por:

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \phi D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\{\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\},$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob estudo (com p parâmetros), avaliado na estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$.

Um valor pequeno da função desvio indica que, para um número menor de parâmetros, obtém-se um ajuste tão bom quanto o ajuste com o modelo saturado. Se denotarmos por $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\hat{\theta}_i^0 = \theta_i(\hat{\mu}_i^0)$ as estimativas de máxima verossimilhança de θ para os modelos

com p parâmetros ($p < n$) e saturado ($p = n$), respectivamente, temos que a função $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ é dada por

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left\{y_i\left(\hat{\theta}_i^0 - \hat{\theta}_i\right) + \left(b(\hat{\theta}_i) - b(\hat{\theta}_i^0)\right)\right\}.$$

Temos, ainda, que $S_p(\mathbf{y};\hat{\boldsymbol{\mu}},\phi) = \frac{1}{\phi}D(\mathbf{y};\hat{\boldsymbol{\mu}})$ é chamada *scaled deviance* e é utilizada quando o parâmetro de dispersão ϕ é diferente de 1.

Uma outra medida de ajuste alternativa é a estatística de Pearson χ^2 generalizada, dada por:

$$\chi^2 = w_i\frac{(y_i - \hat{\mu}_i)^2}{\hat{V}(\hat{\mu}_i)},$$

onde $\hat{V}(\hat{\mu}_i)$ é a função de variância estimada para a distribuição em estudo.

Podemos mostrar que, para respostas com distribuições normais, $S_p(\mathbf{y};\hat{\boldsymbol{\mu}},\phi)$ e $\frac{\chi^2}{\phi}$ tem distribuição χ^2_{n-p} exata e, para respostas com distribuições Binomial e Poisson, tem distribuição χ^2_{n-p} assintótica.

Para verificarmos a adequabilidade do modelo, devemos comparar as medidas $S_p(\mathbf{y};\hat{\boldsymbol{\mu}},\phi)$ e χ^2 com os percentis da distribuição $\chi^2_{n\text{-}p}$. Se $S_p(\mathbf{y};\hat{\boldsymbol{\mu}},\phi) \leq \chi^2_{n\text{-}p,\,\alpha}$ e $\chi^2_{n\text{-}p} \leq \chi^2_{n\text{-}p,\alpha}$ consideramos que há evidências, a um nível de $100\alpha\%$ de probabilidade, que o modelo proposto está bem ajustado aos dados.

Ainda, um valor de $S_p(\mathbf{y};\hat{\boldsymbol{\mu}},\phi)$ próximo de $(n - p)$ pode ser uma indicação de que o modelo ajustado está adequado, já que o valor esperado de uma variável com distribuição χ^2 é dado pela diferença expressa acima.

A análise de uma seqüência de modelos é conhecida como análise da função desvio (ANODEV). Iniciamos pelo modelo mais simples, o modelo nulo, que só tem o intercepto, e, a partir daí, devemos ajustar novos modelos, incluindo mais termos do que nos anteriores, os efeitos de fatores, covariáveis e suas interações.

Tendo ajustado estes modelos, chamados de modelos encaixados, utilizaremos a deviance como uma medida de discrepância do modelo e formaremos uma tabela de diferença de deviances.

Para ilustrar esta análise, suponhamos M_{p1}, M_{p2},..., M_{pr}, uma seqüência de modelos encaixados com a mesma distribuição, a mesma função de ligação e com dimensões,

respectivamente, p_1, p_2,..., p_r. Sejam $\mathbf{X}_{p1}$, $\mathbf{X}_{p2}$,..., $\mathbf{X}_{pr}$, as matrizes dos modelos e $D_{p1} > D_{p2} >...> D_{pr}$, as deviances.

Suponhamos, também, um ensaio inteiramente casualizado, com *r* repetições e dois tratamentos no esquema fatorial, com *a* níveis para o fator *A* e *b* níveis para o fator *B*. Neste caso, obtemos os resultados mostrados na **Tabela 1**.

Tabela 1 – Análise de deviance

Modelo	G.L.	Deviance	Dif. de deviances	Dif. de G.L.	Significado
Nulo	rab - 1	D_1			
			D_1	a - 1	efeito de *A* ignorando *B*
A	$a(rb - 1)$	D_A			
			D_A - D_{A+B}	b - 1	efeito de *B* incluído *A*
A + *B*	$a(rb - 1) - (b - 1)$	D_{A+B}			
			D_{A+B} - D_{A*B}	$(a - 1)(b - 1)$	Interação *AB*, estando incluídos *A* e *B*
A + *B* + *A.B*	$ab(r - 1)$	D_{A*B}			
			D_{A*B}		Resíduo
Saturado	0	0			

Pelos componentes da ANODEV é possível verificarmos a magnitude ou significância dos efeitos desse particular ensaio.

Sejam os modelos M_p e M_q ($p < q$), com p e q parâmetros respectivamente. A estatística $D_p - D_q$, com $(p - q)$ graus de liberdade, é uma medida de variação dos dados explicada pelos termos que estão em M_q e não estão em M_p. Assim, temos que isto equivale a testarmos se os β's são conjuntamente iguais a zero, ou seja, $\beta_{p+1} = \beta_{p+2} = ... = \beta_q = 0$.

Temos, assintoticamente, para ϕ conhecido, que $\frac{1}{\phi}(D_p - D_q) \sim \chi^2_{q-p}$.

Se ϕ é desconhecido, devemos obter uma estimativa consistente $\hat{\phi}$, de preferência baseada no modelo maximal (com *m* parâmetros), e a inferência pode ser baseada na estatística *F*, dada por:

$$F = \frac{(D_p - D_q)/(q - p)}{\hat{\phi}} \sim F_{q-p,n-m}.$$

Para testarmos as hipóteses relativas aos parâmetros β´s apresentamos três estatísticas: ***razão de verossimilhanças***, ***Wald*** e ***escore***.

Seja $\boldsymbol{ß} = \left[\boldsymbol{ß}_1^T, \boldsymbol{ß}_2^T\right]^T$ uma partição do vetor de parâmetros β onde β_1, com dimensão *q* é o vetor de interesse e β_2, com dimensão (*p* – *q*), o vetor *nuisance*, ou seja, um vetor com parâmetros perturbadores, parâmetros que não tenho interesse imediato.

Sejam as hipóteses H_0: $\beta_1 = \beta_{1,0}$ *versus* H_1: $\beta_1 \neq \beta_{1,0}$, onde $\beta_{1,0}$ é um valor especificado para β_1.

Seja $\hat{\boldsymbol{ß}} = \left[\hat{\boldsymbol{ß}}_1^T, \hat{\boldsymbol{ß}}_2^T\right]^T$ o estimador de máxima verossimilhança para β e $\hat{\boldsymbol{ß}}_0 = \left[\hat{\boldsymbol{ß}}_{1,0}^T, \hat{\boldsymbol{ß}}_{2,0}^T\right]^T$, onde $\hat{\boldsymbol{ß}}_{2,0}$ é o estimador de máxima verossimilhança para β_2, sob H_0.

Para testarmos a hipótese H_0, temos:

(1) ***Teste da Razão de verossimilhança***:

$$\Lambda = -2\ln\lambda = 2\left[\ell\left(\hat{\boldsymbol{ß}}_1, \hat{\boldsymbol{ß}}_2; \mathbf{y}\right) - \ell\left(\boldsymbol{ß}_1, \hat{\boldsymbol{ß}}_2; \mathbf{y}\right)\right] = \frac{1}{\phi}\left[D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}})\right],$$

onde $\ell\left(\hat{\boldsymbol{ß}}_1, \hat{\boldsymbol{ß}}_2; \mathbf{y}\right)$ é o logaritmo da função de verossimilhança maximizada sem restrição e $\ell\left(\boldsymbol{ß}_1, \hat{\boldsymbol{ß}}_2; \mathbf{y}\right)$ é o logaritmo da função de verossimilhança maximizada sob H_0.

A regra de decisão é dada por: rejeitamos H_0, a um nível $100\alpha\%$, se $\Lambda > \chi^2_{q,1-\alpha}$.

(2) ***Teste de Wald***: $W = (\hat{\boldsymbol{ß}}_1 - \boldsymbol{ß}_{1,0})^T \left[\hat{Var}(\hat{\boldsymbol{ß}}_1)\right]^{-1} (\hat{\boldsymbol{ß}}_1 - \boldsymbol{ß}_{1,0})$, onde $\hat{Var}(\hat{\boldsymbol{ß}}_1)$ é a $Var(\hat{\boldsymbol{ß}}_1)$ avaliada em $\hat{\boldsymbol{ß}}_0 = \left[\hat{\boldsymbol{ß}}_{1,0}^T, \hat{\boldsymbol{ß}}_{2,0}^T\right]^T$.

A regra de decisão é dada por: rejeitamos H_0, a um nível $100\alpha\%$, se $W > \chi^2_{q,1-\alpha}$.

(3) ***Teste escore***: $E = \mathbf{U}_1^T(\hat{\boldsymbol{ß}}_0)\hat{Var}_0(\hat{\boldsymbol{ß}}_1)\mathbf{U}_1(\hat{\boldsymbol{ß}}_0)$, onde $\hat{Var}_0(\hat{\boldsymbol{ß}}_1)$ é a $Var(\hat{\boldsymbol{ß}}_1)$ avaliada em $\hat{\boldsymbol{ß}}_0 = \left[\hat{\boldsymbol{ß}}_{1,0}^T, \hat{\boldsymbol{ß}}_{2,0}^T\right]^T$.

A regra de decisão é dada por: rejeitamos H_0, a um nível $100\alpha\%$, se $E > \chi^2_{q,1-\alpha}$.

Uma região de confiança assintótica para β_1 obtida a partir da estatística do teste da razão de verossimilhanças, com um coeficiente de confiança de $100(1-\alpha)\%$, é dada por: $2\left[\ell\left(\hat{\boldsymbol{ß}}_1, \hat{\boldsymbol{ß}}_2; \mathbf{y}\right) - \ell\left(\boldsymbol{ß}_1, \hat{\boldsymbol{ß}}_{2,1}; \mathbf{y}\right)\right] < \chi^2_{q,1-\alpha,}$ onde $\hat{\boldsymbol{ß}}_{2,1}$ é a estimativa de máxima verossimilhança para β_2 para cada valor de β_1 que é testado ser pertencente ou não a região.

Pela estatística de *Wald*, temos que uma região com 100(1-α)% de confiança, é expressa por: $W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)^T \left[\hat{Var}(\hat{\boldsymbol{\beta}}_1)\right]^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) < \chi^2_{q,1-\alpha}$.

Até este ponto, vimos que a escolha de um Modelo Linear Generalizado envolve três passos: definição da distribuição, da função de ligação e da matriz do modelo. Para avaliarmos se estas escolhas resultaram em um modelo adequado/apropriado aos dados devemos realizar uma análise de diagnósticos, ou análise dos resíduos.

Nesta fase, faremos a verificação de possíveis afastamentos das suposições feitas, tanto com relação a parte aleatória, a parte sistemática e a função de ligação, bem como a existência de observações extremas (outliers e pontos de alavanca) com alguma interferência desproporcional nos resultados do ajuste.

Uma observação é chamada de outlier da regressão, ou outlier em **y**, se ela se afasta do padrão linear definido pelas outras observações, dando origem a grandes resíduos. Por outro lado, quando uma observação se destaca das demais no espaço das variáveis explanatórias, é chamada de ponto de alavanca. Uma definição de ponto de alavanca é construída fazendo uma analogia entre a solução de máxima verossimilhança para $\hat{\boldsymbol{\beta}}$ num MLG e a solução de mínimos quadrados de uma regressão normal ponderada.

Na convergência do processo iterativo dado pela equação **(3)**, temos que:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{z},$$

onde $\mathbf{z} = \hat{\boldsymbol{\eta}} + \hat{\mathbf{W}}^{1/2}\hat{\mathbf{V}}^{1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}})$. Portanto, $\hat{\boldsymbol{\beta}}$ pode ser interpretado como a solução de mínimos quadrados da regressão linear de $\hat{\mathbf{W}}^{1/2}\mathbf{z}$ contra as colunas $\hat{\mathbf{W}}^{1/2}\mathbf{X}$. A matriz de projeção da solução de mínimos quadrados da regressão linear de **z** contra **X** com pesos **W** fica dada por

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2},$$

que sugere a utilização dos elementos da diagonal principal de $\hat{\mathbf{H}}$ para detectar a presença de pontos de alavanca neste modelo de regressão normal ponderada.

Os tipos de resíduos mais usados para os Modelos Lineares Generalizados são:

a) resíduos ordinários

$$r_i = \mathrm{y}_i - \hat{\mu}_i;$$

b) resíduos de Pearson generalizados

$$r_i^p = \frac{\mathrm{y}_i - \hat{\mu}_i}{\sqrt{\frac{\phi}{w_i} V(\hat{\mu}_i)}},$$

sendo $\hat{\phi}$ uma estimativa consistente do parâmetro ϕ e w_i um peso a priori.

c) resíduos de Pearson generalizados estudentizados internamente

$$r_i^{p'} = \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{\phi}{w_i} V(\hat{\mu}_i)(1-h_i)}}$$

sendo h_i o i-ésimo elemento da diagonal da matriz **H**;

Segundo DEMÉTRIO (2002) há, ainda, os resíduos **componentes da deviance**, **componentes da deviance estudentizado internamente** e **componentes da deviance estudentizado externamente** que podem ser utilizados como medidas de diagnósticos.

Suponhamos, agora, que o logaritmo da função de verossimilhança de β seja definido por $\ell(ß)$. Uma região assintótica de confiança de coeficiente (1 - α) para β é dada por:

$$[ß; 2\{\ell(\hat{ß}) - \ell(ß)\} \leq \chi_p^2(1-\alpha)].$$

Desta forma, para avaliarmos o impacto em $\ell(\hat{ß})$ com a retirada da i-ésima observação podemos utilizar uma medida baseada na região assintótica acima. Esta medida, denominada afastamento da verossimilhança (*likelihood displacement*), é definida por

$$LD_i = 2\{\ell(\hat{ß}) - \ell(\hat{ß}_{(i)})\}.$$

Como não é possível obtermos uma forma analítica para LD_i, utilizamos a segunda aproximação por série de Taylor em torno de $\hat{ß}$, o que nos leva na seguinte expressão:

$$LD_i \cong (ß - \hat{ß})'\{-\ell''(\hat{ß})\}(ß - \hat{ß}).$$

Substituindo $-L''(\hat{ß})$ pelo correspondente valor esperado e β por $\hat{ß}_{(i)}$, obtemos

$$\mathrm{LD_i} \cong \phi(\hat{ß} - \hat{ß}_{(i)})'(\mathbf{X'\hat{W}X})(\hat{ß} - \hat{ß}_{(i)}) \qquad \textbf{(4)}.$$

Assim, temos uma boa aproximação para LD_i quando $\ell(ß)$ for aproximadamente quadrática em torno de $\hat{ß}$.

Como em geral não é possível obtermos uma forma fechada para $\hat{ß}_{(i)}$, utilizamos a aproximação dada por:

$$ß_{(i)}^1 = \hat{ß} - \frac{\hat{r}_i^P \sqrt{w_i \phi^{-1}}}{(1-\hat{h}_{ii})}(\mathbf{X'\hat{W}X})^{-1} x_i,$$

que, na realidade, consiste em tomarmos a primeira iteração do processo iterativo pelo método *scoring* de Fisher quando o mesmo é iniciado em $\hat{ß}$.

Desta forma, substituindo a expressão acima em **(4)**, obtemos:

$$LD_i \cong \left\{ \frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})} \right\} t_{s_i}^2 ,$$

onde $t_{S_i}^2 = \frac{\phi^{1/2}(y_i - \hat{\mu}_i)}{\sqrt{\hat{V}_i(1-\hat{h}_{ii})}}$ é chamado resíduo padronizado.

LD_i é também chamada de ***distância de Cook***.

Algumas técnicas gráficas para verificarmos a adequabilidade do modelo ajustado são apresentadas a seguir:

(a) Resíduos versus alguma função dos valores ajustados: onde utilizamos algum resíduo estudentizado versus $\hat{\eta}$ ou versus os valores ajustados, transformados de tal forma que se tenha uma variância constante para a distribuição utilizada. O padrão nulo deste gráfico é uma distribuição dos resíduos em torno de zero com amplitude constante. Não tem significado para dados binários;

(b) Gráfico da variável adicionada ou da regressão parcial: com este gráfico é possível verificarmos, também, se existe uma relação entre os resíduos do modelo ajustado e uma covariável ainda não incluída no modelo. Para construirmos este gráfico, devemos executar os seguintes procedimentos:

- ajustar um modelo sem a variável a ser adicionada;
- obter deste modelo o resíduo de Pearson generalizado;
- obter v = (I-**H**)$\mathbf{W}^{1/2}\mathbf{u}$, onde **u** é a variável que será adicionada e v representa os resíduos da regressão ponderada de **u** em relação a **X**, com matriz de pesos **W**;
- traçar, então, o gráfico do resíduo de Pearson generalizado versus v.

Se o gráfico apresentar uma certa tendência, ou seja, se os pontos não estiverem espalhados aleatoriamente, a variável **u** é significativa para o modelo;

(c) Gráfico de resíduos parciais ou gráfico de resíduos mais componente: este gráfico é utilizado para verificarmos se uma determinada variável explicativa precisa ser transformada. Inicialmente, ajustamos o modelo com preditor linear $\eta = \mathbf{X}\beta + \gamma\mathbf{u}$, obtendo $\mathbf{W}^{-1}\mathbf{s}$ e $\hat{\gamma}$, sendo **s** o vetor com elementos $s_i = \frac{y_i - \hat{\mu}_i}{a_i(\phi)V(\hat{\mu}_i)} \frac{d\mu_i}{d\eta_i}$. Em seguida, traçamos o gráfico de $\mathbf{W}^{-1}\mathbf{s} + \gamma\mathbf{u}$ versus **u**. Se o gráfico apresentar uma tendência, ou

seja, se os pontos não estiverem espalhados aleatoriamente, a variável **u** é não precisa ser transformada;

(d) Gráfico normal ou *Q-Q plot* (*normal* plot): o gráfico normal de probabilidades nos auxilia na identificação da distribuição originária dos dados e, também, dos valores que se destacam no conjunto.

Para construirmos este gráfico, devemos seguir os seguintes passos:

- ajustar um modelo a um conjunto de dados e obter $d_{(i)}$, os valores ordenados de uma certa estatística de diagnóstico (resíduos, distância de Cook, *h*, etc);
- dada a estatística de ordem na posição *i*, calcular a respectiva probabilidade acumulada p_i e o respectivo quantil, ou seja, o inverso da função de distribuição da variável resposta I_d, no ponto p_i;
- construir o gráfico de $d_{(i)}$ versus I_d.

A ausência da distribuição esperada é verificada quando este gráfico assume as formas:

- **S (esse):** indicando distribuições com caudas muito curtas, isto é, distribuições cujos valores estão muito próximos da média;
- **S invertido (esse invertido):** indicando distribuições com caudas muito longas e, portanto, presença de muitos valores extremos;
- **J e J invertido:** indicando distribuições assimétricas, positivas e negativas, respectivamente.

Estes gráficos são muito dependentes do número de observações e atingem uma estabilidade quando o número de observações é grande.

Um teste formal para verificarmos a adequacidade da função de ligação consiste em adicionarmos $\hat{\boldsymbol{\eta}}^2$ como uma covariável extra e examinarmos a mudança ocorrida na deviance, o que equivale ao teste da razão de verossimilhanças. Se ocorrer uma diminuição drástica, há evidência de que a função de ligação é insatisfatória.

2.1.2 Equações de Estimação Generalizadas

A suposição de independência entre as respostas observadas nem sempre pode ser considerada, como no caso de pesquisas que envolvem o estudo de dados correlacionados.

Dados correlacionados podem surgir a partir de situações tais como:

- dados longitudinais, múltiplas medidas de um mesmo indivíduo em estudo são obtidas ao longo do tempo;

- agrupamentos, as medidas são provenientes de indivíduos que compartilham categorias comuns ou características que conduz a uma correlação. Por exemplo, dados de incidência pulmonar entre os membros de uma família podem estar correlacionados devido a fatores hereditários.

Esta correlação deve ser computada por métodos apropriados de análise de dados. Se a pesquisa em estudo possui dados correlacionados e se eles forem tratados como independentes, teremos como conseqüência:

- inferências incorretas dos parâmetros regressores devido a erros padrões subestimados;
- estimadores ineficientes, ou seja, haverá um maior erro quadrático médio nos estimadores dos parâmetros de regressão

Neste sentido, LIANG & ZEGER (1986) propuseram o uso das Equações de Estimação Generalizadas (EEG), uma extensão dos Modelos Lineares Generalizados, como um método de estimação dos parâmetros do modelo da regressão para tratar dos dados correlacionados.

Os conceitos aqui apresentados podem ser encontrados em HORTON & LIPSITZ (1999) e JOHNSTON (1996).

Para especificarmos um modelo de regressão usando o método EEG, precisamos definir:

- a distribuição da variável resposta;
- a função de ligação;
- as variáveis explicativas;
- a estrutura da matriz de correlação entre observações.

Seja, então, Y_{ij}, $i = 1, 2,....,K$, $j = 1, 2,....,n_i$ representando a j-ésima medida do i-ésimo objeto. Há n_i observações do objeto i e $\sum_{i=1}^{K} n_i$ o total de observações.

Os dados correlacionados são modelados usando a mesma função de ligação e o mesmo preditor linear (componente sistemático) como no caso de independência. O componente aleatório é descrito pela mesma função de variância, mas a estrutura de covariância das medidas correlacionadas também deverão ser modeladas. Seja $Y_i = [Y_{i1},...., Y_{in_i}]'$ o vetor de observações do i-ésimo objeto e $\mu_i = [\mu_{i1},...., \mu_{in_i}]'$ seu correspondente vetor de médias e, seja $\mathbf{V}_i$ o estimador da matriz de covariância de Y_i. A Equação de Estimação Generalizada para estimarmos β é uma extensão da equação de estimação de independência para dados correlacionados e é dada por:

$$\sum_{i=1}^{k} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{ß}} \mathbf{V}_i^{-1} (\mathrm{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{ß})) = 0$$

Seja $\mathbf{R}_i(\alpha)$ uma matriz de correlação "de trabalho" (n_i x n_i) completamente especificada pelo vetor de parâmetros α. A matriz de covariância de Y_i é modelada como:

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\mathbf{a}) \mathbf{A}_i^{\frac{1}{2}},$$

onde:

- $\mathbf{A}_i$ é uma matriz diagonal (n_i x n_i) com $V(\mu_{ij})$, o j-ésimo elemento da diagonal, ou seja, define a variância de Y_{ij} como função da média marginal μ_{ij};

$$\mathbf{A}_i = \begin{pmatrix} V(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & V(\mu_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V(\mu_{in_i}) \end{pmatrix}_{n_i x n_i}$$

- $\mathbf{R}_i(\alpha)$ é a matriz de correlação de trabalho que define a estrutura de dependência entre as medidas repetidas.

Se $\mathbf{R}_i(\alpha)$ é a verdadeira matriz de correlação de trabalho de Y_i, então $\mathbf{V}_i$ será a verdadeira matriz de covariância de Y_i.

A matriz de correlação de trabalho usualmente não é conhecida. Ela é estimada por um processo de ajuste iterativo usando o valor corrente na iteração do vetor de parâmetros β para computar a função apropriada de resíduo de Pearson:

$$r_{ij} = \frac{\mathrm{y}_{ij} - \hat{\mu}_{ij}}{\sqrt{\mathrm{V}(\hat{\mu}_{ij})}}$$

Há diversas estruturas da matriz de correlação de trabalho que podem ser utilizadas para modelar a matriz de correlação de Y_i. Vale salientar que a dimensão do vetor α, que é tratado como um parâmetro de perturbação (*nuisance*), e a forma do estimador de α são diferentes para cada tipo de estrutura. As mais utilizadas são:

(a) Independente ou Identidade: assume independência entre os tempos, isto é,

$$\mathbf{R}_i(\alpha)_{n_i x n_i} = \begin{cases} 1, \text{se } n_i = n_i \\ 0, \text{c.c.} \end{cases} \quad \text{ou} \quad \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Assim EEG = MLG.

(b) Constante no tempo: possui a mesma correlação entre todos os diferentes tempos de avaliação, ou seja,

$$\mathbf{R}_{\mathrm{i}}(\alpha)_{n_i x n_i} = \begin{cases} 1, \text{se } n_i = n_i \\ \alpha, \text{c.c.} \end{cases} \quad \text{ou} \quad \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}, \text{ que pode ser estimado por}$$

$$\hat{\alpha} = \frac{1}{k\phi}\sum_{i=1}^{K}\frac{1}{n_i(n_i-1)}\sum_{j\neq k} r_{ij}r_{ik} \ .$$

(c) Sem estrutura definida: possui diferentes correlações entre todos os tempos de avaliação, ou seja,

$$\mathbf{R}_{\mathrm{i}}(\alpha)_{n_i x n_i} = \begin{cases} 1, \text{se } n_i = n_i \\ \alpha_{n_i x n_i}, \text{c.c.} \end{cases} \quad \text{ou} \quad \begin{pmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1n_i} \\ \alpha_{12} & 1 & \cdots & \alpha_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1n_i} & \alpha_{2n_i} & \cdots & 1 \end{pmatrix}, \text{ que pode ser estimado}$$

por $\hat{\alpha}_{jk} = \frac{1}{k\phi}\sum_{i=1}^{K} r_{ij}r_{ik}$.

Número de parâmetros a serem estimados: $n_i(n_i\text{-}1)\frac{1}{2}$

(d) Auto-regressiva: assume uma relação entre as correlações referentes aos tempos anterior e posterior, ou seja,

$$\mathbf{R}_{\mathrm{i}}(\alpha)_{n_i x n_i} = \begin{cases} 1, \text{se } n_i = n_i \\ \alpha^{|n_i - n_i|}, \text{c.c.} \end{cases} \quad \text{ou} \quad \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{n_i-1} \\ \alpha & 1 & \cdots & \alpha^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \cdots & 1 \end{pmatrix}, \text{ que pode ser}$$

estimado por $\hat{\alpha} = \frac{1}{k\phi}\sum_{i=1}^{K}\frac{1}{n_i-1}\sum_{j\leq n_i-1} r_{ij}r_{i,j+1}$.

(e) M-dependente: depende das M observações anteriores, ou seja,

$$\mathbf{R}_i(\alpha)_{n_i x n_i} = \begin{cases} 1, \text{se } n_i = n_i \\ \alpha_t, \text{c.c.} \end{cases} \quad \text{ou} \quad \begin{pmatrix} 1 & \alpha_1 & \cdots & \alpha_{n_i-1} \\ \alpha_1 & 1 & \cdots & \alpha_{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n_i-1} & \alpha_{n_i-2} & \cdots & 1 \end{pmatrix}, \text{ que pode ser}$$

estimado por $\hat{\alpha}_t = \frac{1}{k\phi}\sum_{i=1}^{K}\frac{1}{n_i - t}\sum_{j \le n_i - t} r_{ij} r_{i,j+t}$

Número de parâmetros a serem estimados: $0 < M \le n_i - 1$

(f) Fixada: as correlações são fixadas pelo pesquisador, ou seja,

$$\mathbf{R}_i(\alpha)_{n_i x n_i} = \begin{cases} 1, \text{se } n_i = n_i \\ \mathrm{r}_{n_i x n_i}, \text{c.c.} \end{cases} \quad \text{ou} \quad \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n_i} \\ r_{12} & 1 & \cdots & r_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n_i} & r_{2n_i} & \cdots & 1 \end{pmatrix}$$

O ajuste de um modelo específico utilizando EEG segue os seguintes passos:

- calcular o estimador inicial de β, por exemplo com um Modelo Linear Generalizado, assumindo independência;
- computar a matriz de correlação de trabalho $\mathbf{R}_i(\alpha)$;
- calcular o estimador da covariância

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \hat{\mathbf{R}}_i(\alpha) \mathbf{A}_i^{\frac{1}{2}}$$

- atualizar β:

$$\beta_{r+1} = \beta_r - \left[\sum_{I=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial ß}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial ß}\right]^{-1} \left[\sum_{I=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial ß}' \mathbf{V}_i^{-1} (\mathrm{Y}_i - \boldsymbol{\mu}_i)\right]$$

- repetir até convergir.

As EEG tem como boa propriedade estatística:

- $\sqrt{K}(\hat{ß} - ß) \to N_k(0, \mathbf{M}(\phi))$ se o modelo médio está correto, ainda que $\mathbf{V}_i$ esteja especificada incorretamente, onde:
  - $\mathbf{M}(\phi) = \mathbf{I}_0^{-1} \mathbf{I}_1 \mathbf{I}_0^{-1}$
  - $\mathbf{I}_o = \sum_{I=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial ß}' \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial ß}$
  - $\mathbf{I}_1 = \sum_{I=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial ß}' \mathbf{V}_i^{-1} Cov(\mathrm{Y}_i) \frac{\partial \boldsymbol{\mu}_i}{\partial ß} \mathbf{V}_i^{-1}$

Esta propriedade significa que não precisamos especificar corretamente a matriz de correlação para termos um estimador consistente para os parâmetros regressores. Escolher a correlação de trabalho mais próxima da verdadeira correlação, aumenta a eficiência estatística dos estimadores dos parâmetros da regressão, assim, devemos especificar a matriz de correlação tão precisa quanto o possível baseada em conhecimentos específicos da área em estudo.

A $Cov(\hat{\boldsymbol{\beta}})$ é dada por:

$$Cov(\hat{\boldsymbol{\beta}}) = \mathbf{I}_0^{-1}.$$

Esta é a inversa da Matriz de Informação de Fisher, muitas vezes utilizada em Modelos Lineares Generalizados como um estimador da covariância do estimador de máxima verossimilhança de β. É um estimador consistente da matriz de covariância de $\hat{\beta}$ se o modelo médio e a matriz de correlação forem especificadas corretamente.

O estimador da matriz de covariância de $\hat{\beta}$, dado por:

$$\mathbf{M} = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$$

é chamado empírico, ou robusto, pois, tem a propriedade de ser um estimador consistente, até mesmo se a matriz de correlação está mal especificada, ou seja, se $\text{Cov}(Y_i) \neq \mathbf{V}_i$.

No cálculo de $\mathbf{M}$, β e ϕ são substituídos por suas estimativas iniciais e, a $\text{Cov}(Y_i)$ é substituída por uma estimativa, como:

$$(Y_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}))'(Y_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})).$$

2.1.3 Seleção de Modelos

Uma das etapas de grande importância da construção de um modelo de regressão é a seleção do modelo. Considerando que temos um conjunto de variáveis que seriam possíveis candidatas a variáveis explicativas em um modelo, vamos comentar alguns procedimentos para obter um modelo parcimonioso e que se ajuste bem aos dados.

Um primeiro procedimento, conhecido como de todas as regressões possíveis, consiste em fazermos todos os ajustes possíveis. Por exemplo, com 4 variáveis explicativas, devemos ajustar $(2^4\text{-}1)$ modelos. Para cada modelo ajustado, podemos calcular um ou mais critérios para compará-los: R^2_p, MSE_p, C_p, $PRESS_p$ e AIC.

Na utilização deste procedimento, vamos levar em consideração algumas suposições:

(a) o número de potenciais variáveis é P – 1;

(b) todos os modelos tem intercepto: β_0;

(c) p é o número de variáveis em um particular modelo, $1 \leq p \leq P$;

(d) o número de observações, n, é maior que o número máximo de variáveis explicativas, $n > p$, $n >$ P.

Os quatro primeiros critérios descritos a seguir (R^2_p, $R^2_{a,p}$, C_p e $PRESS_p$) e aplicados em MLG foram citados e expressos em notas de aula por BARRETO (2005). O critério AIC foi citado como critério de seleção em PAULA (2004).

I. SSE_p ou R^2_p (coeficiente de determinação)

Corresponde ao cálculo do coeficiente de determinação do modelo ajustado, dado por:

$$R^2_p = 1 - \frac{D_p}{D_\mu},$$

em que D_p é a deviance do modelo sob estudo e D_μ é a deviance do modelo com um único parâmetro. Esta quantidade mede a redução da variação total em **Y** ao utilizarmos o modelo com as variáveis em questão e tem as seguintes propriedades:

- $0 \leq R^2_p \leq 1$;
- valores grandes de R^2_p, em geral, indicam que o modelo está bem ajustado. Assim, valores próximos de zero, indicam um modelo pobre ou um ajuste ruim;
- em geral, valores altos de b_i produzem R^2_p grande, sem, no entanto, o modelo estar bem ajustado. Assim, nem sempre é bom usar só R^2_p;
- quanto maior o número de variáveis explicativas, maior será o valor de R^2_p. Para comparar R^2_p de modelos com um número diferente de variáveis explicativas, devemos utilizar o critério $R^2_{a,p}$

II. MSE_p ou $R^2_{a,p}$ (coeficiente de determinação ajustado)

Este critério, ao contrário do anterior, leva em consideração o número de parâmetros do modelo e dado por:

$$R^2_{a,p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{D_p}{D_\mu},$$

onde D_p é a deviance do modelo sob estudo e D_μ é a deviance do modelo com um único parâmetro.

III. C_p

Este critério é expresso como:

$$C_p = \frac{D_p}{\frac{D_{p-1}}{(n-(p-1))}} - (n-2p)$$

onde D_p é a deviance do modelo sob estudo e D_{p-1} é a deviance do modelo sem a variável explicativa que está sendo adicionada no modelo com D_p.

Usando este critério é possível identificarmos o subconjunto de variáveis para os quais:

(a) o valor de C_p é pequeno;

(b) o valor de C_p é próximo de p, o número de parâmetros do modelo sob estudo.

Os subconjuntos com valores pequenos de C_p tem uma deviance pequena, enquanto que o valor de C_p próximo de p, indica que o vício do modelo de regressão é pequeno.

IV. PRESS$_p$ (soma de quadrados dos valores preditos)

Este critério é uma medida do quão bom é usarmos os valores preditos do particular modelo para podermos prever as respostas observadas, indicadas por y_i.

$$\text{PRESS}_p = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2 ,$$

onde $\hat{y}_{i(i)}$ é o valor predito da observação quando ela foi omitida do ajuste.

Modelos com valores pequenos de PRESS$_p$ são considerados bons modelos.

V. **Critério de Akaike**

Neste caso, o critério utilizado consiste em encontrarmos o modelo tal que a quantidade abaixo seja minimizada

$$\text{AIC} = D_p + 2p,$$

em que D_p denota a deviance do modelo e p o número de parâmetros.

A partir da obtenção dos cinco critérios para cada modelo, é possível, então, identificarmos aqueles que seriam candidatos a “melhores modelos”.

Em caso de empates, devemos optar pelo modelo com menor número de variáveis explicativas (princípio da parcimônia).

Um segundo procedimento de seleção de modelos é conhecido como -procedimento de seleção de variáveis stepwise. Ele consiste em um algoritmo que pode adicionar ou remover variáveis, tendo como critério a variação da deviance, o coeficiente de correlação parcial, a estatística de teste do parâmetro associado, a estatística F* ou o valor p.

Seu uso é indicado quando o número de variáveis explicativas é grande.

A diferença deste método com o anterior é que, aqui, apenas um critério é utilizado.

Há três maneiras de se utilizar o algoritmo stepwise:

(I) Método forward

Iniciamos o método pelo modelo mais simples que contém apenas o intercepto, isto é, $\mu = \alpha$. Ajustamos, então, para cada variável explicativa o modelo:

$$\mu = \alpha + \beta_j x_j, \ \left(j = 1,..., \ p\right).$$

Testamos H_0: $\beta_j = 0$ contra H_1: $\beta_j \neq 0$. Seja P_E o nível crítico do teste especificado a priori e P o menor nível descritivo entre os p testes. Se $P \leq P_E$, a variável correspondente entra no modelo. Suponhamos que X_1 tenha sido escolhida. Então, no passo seguinte, ajustamos os modelos

$$\mu = \alpha + \beta_1 x_1 + \beta_j x_j, \ \left(j = 2,..., \ p\right).$$

Testamos H_0: $\beta_j = 0$ contra H_1: $\beta_j \neq 0$. Seja P o menor nível descritivo entre os $(p - 1)$ testes. Se $P \leq P_E$, a variável correspondente entra no modelo. Repetimos o procedimento até que ocorra $P > P_E$.

(II) Método backward

Iniciamos o método pelo modelo

$$\mu = \alpha + \beta_1 x_1 + ... + \beta_p x_p .$$

Testamos H_0: $\beta_j = 0$ contra H_1: $\beta_j \neq 0$, para $j = 1,..., p$. Seja P o maior nível descritivo entre os p testes. Se $P > P_S$, a variável correspondente sai do modelo. Suponhamos que X_1 tenha saído do modelo. Então, no passo seguinte, ajustamos o modelo

$$\mu = \alpha + \beta_2 x_2 + ... + \beta_p x_p .$$

Testamos H_0: $\beta_j = 0$ contra H_1: $\beta_j \neq 0$, para $j = 1,..., p$. Seja P o maior nível descritivo entre os $(p - 1)$ testes. Se $P > P_S$, a variável correspondente sai do modelo. Repetimos o procedimento até que ocorra $P \leq P_S$.

(III) Método stepwise

É uma mistura dos dois métodos acima. Iniciamos o processo com o modelo $\mu = \alpha$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira não sai do modelo. O processo continua até que nenhuma variável seja incluída ou seja retirada do modelo.

2.1.4 Geoestatística

A metodologia Geoestatística foi formalizada pelo engenheiro Georges Matheron, na França, final da década de 60, a partir de estudos práticos do cálculo de reservas de minas de ouro na África do Sul desenvolvidos por vários pesquisadores, onde se destacaram o engenheiro de minas Daniel G. Krige e o estatístico H. S. Sichel.

Esta metodologia se preocupa com o estudo das variáveis regionalizadas ou variáveis com condicionamento espacial, e tem como idéia básica o fato de que quanto mais próximos estiverem dois pontos amostrados, espera-se que seus valores sejam semelhantes.

O valor de uma amostra localizada espacialmente em x_1, onde x_1 é um conjunto de coordenadas geográficas, é interpretado como uma realização $y(x_1)$ da variável regionalizada $Y(x_1)$. Em um espaço ou região de estudo no qual se dispersa o conjunto de amostras, temos as realizações das n variáveis regionalizadas $Y(x_1)$, $Y(x_2)$,..., $Y(x_n)$ correlacionadas entre si. Conforme veremos, uma variável regionalizada que depende, então, de sua posição espacial, será utilizada para medir a variação espacial de um fenômeno em estudo.

A variável regionalizada é contínua no espaço já que possui como propriedade o fato de apresentar valores muito próximos em dois pontos vizinhos e progressivamente mais diferentes a medida que os pontos vão se distanciando. Apesar disto, não é possível conhecermos os seus valores em todos os pontos, mas sim apenas em alguns que foram obtidos por amostragem.

A Geoestatística tem como principal objetivo analisar os valores de uma variável distribuída no espaço para se determinar sua estrutura de dependência espacial e, assim, efetuar interpolações. Ao extrair dos dados disponíveis uma imagem da sua variabilidade e uma medida da correlação existente entre valores tomados em dois pontos do espaço, temos uma análise estrutural ou análise da estrutura no espaço. A estimativa de dependência entre amostras é feita através do semivariograma. Já a previsão ou interpolação, isto é, inferência sobre a realização do processo em localizações não medidas, é feita através de um interpolador geoestatístico chamado de krigagem, em homenagem ao engenheiro de minas D. G. Krige.

Na etapa da análise estrutural, veremos que há dois tipos importantes de estrutura espacial: estacionariedade e isotropia. Ao admitirmos a hipótese de estacionariedade, assumimos que o processo é similar ao longo da região de estudo. Ainda, se o processo é estacionário e isotrópico, a estrutura de pequena escala depende das localizações espaciais apenas através da distância euclideana entre elas, ou seja, é invariante sob rotação e translação das localizações.

Resumidamente, os passos de um estudo por técnicas geoestatísticas são:

(a) análise exploratória dos dados;

(b) análise estrutural (cálculo do semivariograma experimental e ajuste do modelo teórico) e

(c) previsão (interpolação) (krigagem e simulação).

Os conceitos aqui apresentados podem ser encontrados em CAMARGO (1997), CAMARGO et al. (2001), DIGGLE et al. (2000), JIAN et al. (1996), LANDIM (2003), MELLO et al. (2005) e SOARES (2006).

A região de estudo que contém um número discreto de amostras será denotada por *A*.

O formato básico de dados geoestatísticos univariados é dado por:

$$(x_i, y_i),\ i = 1, ..., n,$$

onde x_i identifica a localização espacial (geralmente em duas dimensões, embora uma e três dimensões possam ocorrer); y_i é uma medida escalar tomada na posição x_i.

O modelo geoestatístico é um processo estocástico $\{Y(x): x \in A\}$, o qual é considerado ser uma realização parcial do processo estocástico $\{Y(x): x \in R^2\}$.

O processo estocástico $Y(x)$ é Gaussiano se a distribuição conjunta de $Y(x_1)$,...., $Y(x_n)$ é Gaussiana Multivariada para qualquer inteiro n e um conjunto de localizações x_i.

Seja $Y(x)$ uma função aleatória do conjunto de variáveis aleatórias $Y(x_1)$, $Y(x_2)$,..., $Y(x_n)$, correlacionadas entre si e localizadas espacialmente em A, e $\{Y(x): x \in R^2\}$ um processo espacial Gaussiano, com as seguintes propriedades:

- a função média: $\mu(x) = E[Y(x)]$, muitas vezes chamada de tendência;
- a função covariância: Cov $\{Y(x), Y(x')\}$, onde $x' = x + u$ e u uma determinada distância;
- e, a função variância: $\sigma^2(x) = \text{Var}\{Y(x)\}$.

Além disso, o processo $Y(x)$ é estacionário se:

- $E[Y(x)] = \mu$ for constante para todo x, ou seja, $E[Y(x_1)] = E[Y(x_2)] = ... = E[Y(x_n)] = \mu$;
- $\text{Cov}\{Y(x), Y(x')\} = \text{Cov}(x' - x)$ depende somente da diferença entre duas localizações de interesse.

A hipótese de estacionariedade em relação à covariância é, então, definida considerando que a correlação entre duas variáveis aleatórias depende somente da distância espacial que as separa e é independente da sua localização. Esta hipótese de que a correlação entre quaisquer duas variáveis aleatórias distanciadas espacialmente de um vetor u depende somente de u, implica que a covariância pode ser expressa como:

$$Cov\{Y(x),Y(x')\} = Cov\{Y(x),Y(x+u)\} =$$
$$= C(u) = \rho(u)\sqrt{Var[Y(x)]Var[Y(x')]},$$

onde $u = x' - x$ e $\rho(u)$ é a função de correlação dada por:

$$\rho(Y(x),Y(x')) = \rho(Y(x),Y(x+u)) = \rho(u) = \frac{C(u)}{\sqrt{Var[Y(x)]Var[Y(x')]}} \ .$$

Admitindo a estacionariedade em relação à esperança, ou seja, que a média da variável regionalizada no ponto (x) é igual àquela no ponto ($x + u$), temos que $E[Y(x)] = E[Y(x')]$ e, conseqüentemente, $E[Y^2(x)] = E[Y^2(x')]$. Desta forma, para o processo estacionário, a variância de $Y(x)$ é constante e é útil para escrever a função covariância como $C(u) = \sigma^2\rho(u)$ (Apêndice B – B1), onde σ^2 é a variância, $\rho(.)$ é a função correlação e u é a distância que separa as duas variáveis aleatórias.

O processo é estacionário e isotrópico se, adicionalmente, a covariância depende somente da distância que separa as duas variáveis aleatórias, de modo que $Cov(x' - x) = Cov(||x' - x||) = C(||x' - x||)$, onde $||\ .\ ||$ denota a distância Euclideana.

Uma relação próxima à função covariância é o variograma, uma ferramenta básica que nos permite descrever quantitativamente a variação de um fenômeno regionalizado no espaço (HUIJBREGTS, 1975). Assim, a estimativa da dependência entre os pontos amostrais vizinhos no espaço pode ser realizada através da autocorrelação e o variograma é o instrumento mais indicado na estimativa desta dependência. Assim como a função covariância, o variograma também é uma medida média da correlação entre duas variáveis aleatórias. A estrutura de correlação espacial de um conjunto de dados é, então, definida a partir da comparação de valores tomados simultaneamente em dois pontos, segundo uma determinada direção. A função variograma $2\gamma(x, x')$ é definida como sendo a esperança matemática do quadrado da diferença entre os valores de pontos no espaço, separados por uma distância u, conforme a seguinte equação:

$$2\gamma(u) = E\{[Y(x+u) - Y(x)]^2\},$$

e, através de uma amostra $y(x_i)$, $i = 1, ..., n$, pode ser estimada por:

$$2\gamma^*(u) = \frac{1}{N(u)}\sum_{i=1}^{N}[y(x+u) - y(x)]^2 \ ,$$

onde $N(u)$ é o número de pares de pontos separados por uma distância u, $y(x)$ é o valor da variável regionalizada no ponto x e $y(x + u)$ é o valor da variável regionalizada no ponto $(x + u)$.

A função variograma $2\gamma(u)$ pode ser expressa em termos da variância σ^2 e da função covariância $C(x, x') = C(x, x+ u) = C(u)$:

$$2\gamma(u) = 2\sigma^2 - 2C(u).$$ (Apêndice B – B2)

Para um processo com estrutura de covariância estacionária, o variograma se reduz a:

$$\gamma(u) = \frac{1}{2}\{2\sigma^2 - 2C(u)\} = \sigma^2 - C(u) = \sigma^2\{1 - \rho(u)\}.$$ (Apêndice B – B2)

A função $\gamma(u)$ é denominada função semivariograma, que é a metade da função variograma. Esta função $\gamma(u)$ é vetorial, já que é uma função do vetor *u* e, portanto depende da magnitude e direção de *u.* Vale salientar que o vetor *u* é composto por uma componente direção (α), bem como o valor da distância. Estas considerações nos revelam que é possível construirmos semivariogramas segundo uma direção predefinida (norte-sul (N-S): α = 0º, nordeste-sudoeste (NE-SO): α = 45º, leste-oeste (L-O): α = 90º e noroeste-sudeste (NO-SE): α = 135º - **Figura 1**). Na prática, construímos semivariogramas segundo várias direções para conhecermos a estrutura da variável regionalizada em estudo, ou melhor, para verificarmos se há uma diferença na estrutura dos dados ao longo das direções. Quando os semivariogramas mostram diferentes comportamentos para diferentes direções ocorre o que chamamos de anisotropia e, assim, a distribuição espacial do fenômeno não é isotrópica, ou seja, sem direções privilegiadas de variabilidade.

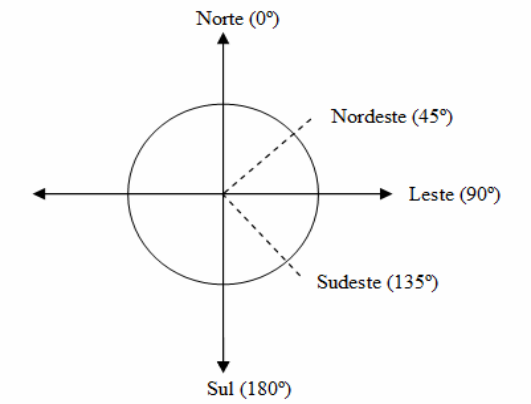


Figura 1. Convenções direcionais usadas na Geoestatística

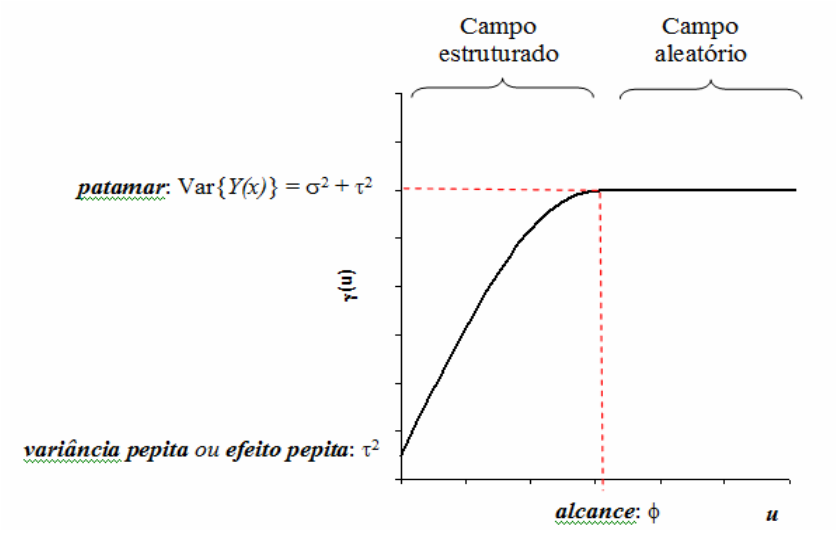
O procedimento de obtenção e análise do semivariograma é chamado de análise estrutural.

A interpretação do semivariograma nos permite descrever o comportamento espacial das variáveis regionalizadas através dos seus parâmetros estruturais:

- variância pepita ou efeito pepita *(*"nugget"*)*: τ^2*;*
- patamar ("sill"): Var{*Y(x)*} = $\sigma^2 + \tau^2$;
- alcance, amplitude ou raio ("range"): ϕ.

Pelo semivariograma conseguimos identificar qualquer valor de $Y(x)$ correlacionado com outros valores $Y(x + u)$ que estiverem dentro de um raio ϕ (denominado alcance ou amplitude) de x. Esta correlação, ou melhor, a influência de um valor em outro, decresce conforme $Y(x + u)$ aproxima-se de ϕ.

Graficamente, podemos ilustrar o semivariograma como segue:



- **alcance ou amplitude** (ϕ)**:** é a distância a partir da qual as amostras passam a ser independentes. Em outras palavras, a amplitude reflete o grau de homogeneização entre as amostras, ou seja, quanto maior for a amplitude maior será a homogeneidade entre as amostras. Nesse sentido, o semivariograma dá um significado preciso da noção tradicional de zona de influência. A amplitude (ϕ) é a distância que separa o campo estruturado (amostras correlacionadas) do campo aleatório (amostras independentes);
- **patamar** ($\text{Var}\{Y(x)\} = \sigma^2 + \tau^2$)**:** é o valor no qual o semivariograma estabiliza-se (no campo aleatório), é o ponto a partir do qual as amostras tornam-se independentes devido à grande distância que as separa;
- **efeito pepita** (τ^2)**:** é o valor da função semivariograma na origem ($u = 0$). Teoricamente esse valor deveria ser zero, pois duas amostras tomadas no mesmo ponto ($u = 0$) deveriam ter os mesmos valores; entretanto quando não é assim, atribui-se, esta diferença, geralmente, a erros de amostragem e/ou análise devido à variabilidade natural da localização de amostragem;

Em um processo estacionário, a função de covariância depende de um argumento escalar u e da definição da função de correlação $\rho(.)$, o que implicaria que $\rho(0) = 1$, desde que

$Y(x)$ seja perfeitamente correlacionado com ele mesmo. Entretanto, na prática, é razoável investigar que o valor medido na localização x poderia estar replicado e que o resultado de múltiplos valores não seriam idênticos. Matematicamente, isto implica que $\rho(0) < 1$.

Um modelo físico para corresponder a este comportamento seria que o $Y(x)$ inclui um componente da variação aleatória devido ao erro de medida e, a representação estatística dos valores $Y_1,....,Y_n$, provenientes de localizações não necessariamente distintas $x_1,....,x_n$, segue o modelo:

$$Y_i = S(x_i) + \varepsilon_i,\ i = 1,..., n, \qquad \textbf{(5)}$$

na qual $S(x)$ é um processo Gaussiano estacionário com função de covariância $C_s(u) = \sigma^2\rho(u)$, tal que $\rho(0) = 1$ e ε_i são variáveis aleatórias mutuamente independentes com distribuição $N(0, \tau^2)$;

Seja, então, o processo $Y(x)$ na qual o valor da localização x é obtido pelo modelo estatístico dado por equação **(5)**, com $x = x_i$. Então, $Y(x)$ tem variância $\sigma^2 + \tau^2$ e função de covariância $C(u) = \sigma^2\rho(u)$ e, daqui, a função de correlação:

$$\rho_Y(u) = \sigma^2\rho(u) / (\sigma^2 + \tau^2) \rightarrow \sigma^2 / (\sigma^2 + \tau^2) < 1,$$

com $u \rightarrow 0$.

A correspondente função semivariograma deste processo é expressa por:

$$\gamma_Y(u) = \tau^2 + \sigma^2\{1 - \rho(u)\} = \tau^2 + \gamma_s(u),$$

onde o parâmetro τ^2 é a variância pepita, $\text{Var}\{Y(x)\} = \sigma^2 + \tau^2$ é o patamar, $\rho(u)$ é a função de correlação que depende do parâmetro ϕ, denominado alcance ou amplitude.

Este resultado nos mostra que o efeito do termo do erro de medida em **(5)** está presente para introduzir um intercepto diferente de zero para o semivariograma.

Através do efeito pepita e patamar conseguimos determinar o grau de aleatoriedade presente nos dados através da seguinte expressão E = efeito pepita/patamar = $\tau^2/(\sigma^2 + \tau^2)$, na qual:

$E < 0,15$: componente aleatória pequena;

$0,15 \leq E \leq 0,30$: componente aleatória significativa;

$E \geq 0,30$: componente aleatória muito significativa.

Para $E \geq 0,30$ temos um modelo de pepita pura, no qual não ocorre covariância entre os valores e, desta forma, a análise semivariográfica não se aplica, nos revelando, assim, que devemos utilizar outros métodos de interpolação.

O procedimento a ser executado para o cálculo do semivariograma depende do tipo de amostragem realizada, sendo mais comum trabalharmos com pontos amostrais irregularmente distribuídos no espaço.

Considere o conjunto de pontos amostrados e regularmente espaçados em duas dimensões (x, y), com distância *u* unidades de medida (u.m.) entre dois pontos consecutivos, conforme ilustrado na **Figura 2**.

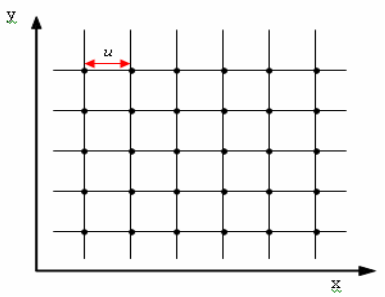


Figura 2. Amostras regularmente espaçadas em duas dimensões

O cálculo do semivariograma em uma determinada direção α deverá ser efetuado para as todas as distâncias (u, 2u,...). Suponhamos uma direção $\alpha = 90^o$, então, incluímos no cálculo de $\gamma^*(\alpha = 90^o, u)$ todos os pares de pontos amostrais na direção Leste distantes u metros. Em seguida, para todos os pares amostrais distantes 2u m, calculamos $\gamma^*(\alpha = 90^o, 2u)$ e, assim, sucessivamente, o processo é repetido para todas as distâncias, dada uma direção α específica.

Quando temos uma amostragem irregularmente distribuída no espaço bidimensional (x, y) torna-se impossível, de inicio, encontrarmos pares de amostras suficientes com exatamente a mesma distância u para o cálculo do semivariograma em uma determinada direção. Para evitarmos este problema, devemos definir uma distância de tolerância Δu para o espaçamento u entre os pares de amostras e um ângulo de tolerância $\Delta\alpha$ para a direção α considerada. Desta forma, para o cálculo do semivariograma de uma distribuição irregular de pontos ao longo de uma direção α, consideramos todas as amostras que se encontram no ângulo $\alpha+\Delta\alpha$ e, em seguida, classificamos os pares de amostras em classes de distância $u+\Delta u$, $2u+\Delta u$, ..., onde u é a distância básica. Este esquema de seleção dos pares de pontos amostrais está ilustrado em **Figura 3**.

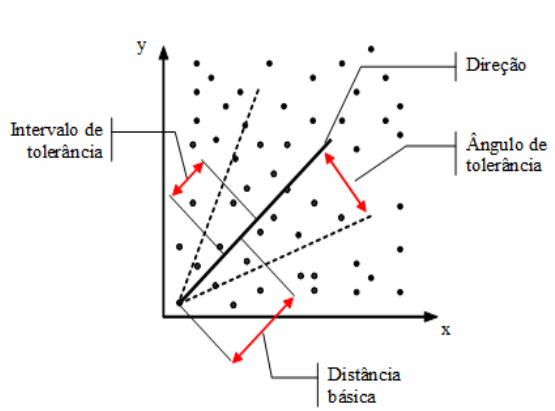


Figura 3. Esquema de obtenção de valores para o semivariograma a partir de uma grade irregular.

Após gerarmos o semivariograma experimental, é necessário ajustarmos uma função matemática que expresse a estrutura de dependência espacial da variável regionalizada em estudo. Este ajuste de uma função matemática é conhecido como ajuste de modelos teóricos ao semivariograma experimental.

O semivariograma como ferramenta básica será utilizado para calcularmos os valores da semivariância para uma dada distância, os quais são necessários para a organização do sistema de equações da krigagem (interpolação geoestatística). O semivariograma de pontos, chamado de semivariograma experimental, não serve para este fim, porque há necessidade de interpolação e, invariavelmente, os pontos se apresentarão com uma certa dispersão, principalmente para distâncias grandes, quando o número de pares de amostras vai diminuindo. Devido a este fato, devemos ajustar uma função matemática que descreva continuadamente a variabilidade ou correlação espacial existente nos dados.

O ajuste de modelos teóricos ao semivariograma experimental é feito de maneira interativa, onde, a partir dos parâmetros do semivariograma (modelo, efeito pepita, amplitude e patamar), o semivariograma teórico é desenhado juntamente como os pontos do semivariograma experimental e, se o ajuste não for satisfatório, novos parâmetros são fornecidos e assim sucessivamente até que o ajuste seja considerado satisfatório.

A maioria dos modelos paramétricos do semivariograma utilizados na prática incluirão um efeito pepita e, no caso estacionário, são dados por:

$$\gamma(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}.$$

Há diversos tipos de funções de correlação ($\rho(u)$) que podem ser utilizadas sob a condição de positividade, ou seja, $\rho(u)$ deve ser uma função positiva-definida, que incorpore as seguintes características:

- $\rho(.)$ é monótona não crescente em u (a correlação entre duas medidas decai com o incremento da distância entre duas localizações amostrais);
- $\rho(u) \rightarrow 0$ quando $u \rightarrow \infty$ (a correlação decai para zero para grandes distâncias de separação);
- pelo menos um parâmetro do modelo controla a taxa na qual $\rho(u)$ decai para zero (já que a distância de separação na qual a correlação torna-se insignificante não será conhecida antecipadamente).

Dadas as considerações acima, algumas famílias candidatas à função de correlação que normalmente cobrem a generalidade das situações de dispersão de fenômenos espaciais são: família Esférica, Exponencial Potência e Matérn.

A família de correlação esférica depende de um único parâmetro de escala ϕ e é definida por:

$$\rho(|u|;\phi) = \begin{cases} 1 - \frac{3}{2}\left(\frac{|u|}{\phi}\right) + \frac{1}{2}\left(\frac{|u|}{\phi}\right)^3 = 1 - \left(\frac{3}{2}\left(\frac{|u|}{\phi}\right) - \frac{1}{2}\left(\frac{|u|}{\phi}\right)^3\right) = 1 - (\text{Sph}(|u|)), & 0 \leq |u| \leq \phi \\ 0 & , \quad |u| > \phi \end{cases}$$

A família exponencial potência com dois parâmetros, k e ϕ, é definida por:

$$\rho(|u|) = \exp\{-(|u|/\phi)^k\} = Pot(|u|) \ ,$$

com $\phi > 0$ e $0 < k \leq 2$;

Quando $k = 1$, temos a família de função de correlação Exponencial, Exp($|u|$), e quando $k = 2$ é chamada de função de correlação Gaussiana, Gau($|u|$).

A família Matérn é expressa por:

$$\rho(u;\phi,k) = \{2^{k-1}\Gamma(k)\}^{-1}(u/\phi)^k K_k(u/\phi) \, ,$$

onde (ϕ, k) são parâmetros e $K_k(.)$ é a função Bessel modificada de terceiro tipo de ordem k.

Esta família é valida para $\phi > 0$ e $k > 0$. Se $k = 0.5$, temos uma função de correlação Exponencial, $\rho(u) = \exp(-u/\phi)$.

Para que o modelo espacial apresente um bom poder preditivo é necessário inicialmente identificarmos se a variável em estudo apresenta uma direção privilegiada de variabilidade, isto é, se existe algum efeito externo que faz com que a variável se espalhe mais intensamente em uma determinada direção. Neste caso, devemos incorporar esta tendência no modelo teórico do semivariograma buscando representar de forma mais adequada, ou mais próxima possível da realidade, a variabilidade espacial do fenômeno em estudo.

As condições ambientais, por exemplo, podem induzir estes efeitos direcionais (vento, formação do solo, etc) e, como conseqüência, a correlação espacial pode variar com a direção. Quando isto é verificado, dizemos que a distribuição espacial do fenômeno em estudo, por apresentar um espalhamento mais intenso em uma determinada direção, é denominada anisotrópica.

Neste sentido, para uma correta inferência sobre a realização do processo em localizações não medidas, ao modelarmos a estrutura de correlação espacial devemos levar em consideração a existência da anisotropia, incorporando as direções de maior e menor espalhamento da variável regionalizada em estudo.

A anisotropia pode ser constatada de diversas maneiras, como por exemplo, através da observação dos semivariogramas obtidos para as diferentes direções convencionadas na

Geoestatística: 0^o, 45^o, 90^o e 135^o. Se os semivariogramas são distintos nestas direções, o modelo é denominado anisotrópico. Se do contrário, ou seja, se o semivariograma apresenta uma forma semelhante em todas as direções no espaço, a estrutura do fenômeno é denominada isotrópica, sem direções privilegiadas de variabilidade.

Outra forma de detectá-la é através do esboço de um gráfico de uma elipse, também denominado diagrama da rosa, calculado através dos alcances obtidos em direções distintas.

Outra opção, considerada por CAMARGO (1997) como a forma mais direta e eficiente para se identificar a presença da anisotropia, é a utilização do mapa de semivariograma, ou semivariograma de superfície, um gráfico em 2D que ilustra a variabilidade espacial do fenômeno em estudo e que nos permite identificar os eixos de anisotropia, onde o eixo maior da elipse corresponderá a direção de maior variabilidade do fenômeno nos indicando, assim, o ângulo de anisotropia e, o eixo menor (ortogonal ao maior) ilustrará a direção de menor variabilidade.

A forma mais comum de anisotropia segue o comportamento de anisotropia geométrica, **Figura 4**, onde observamos um patamar constante e alcances variando conforme as direções.

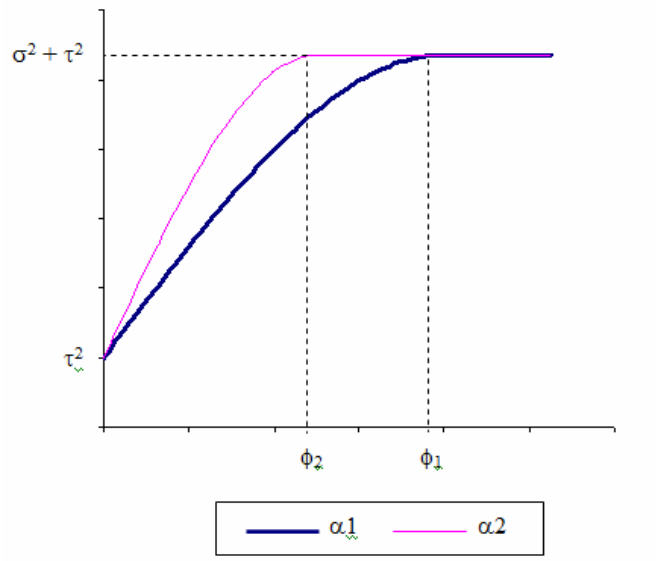


Figura 4. Anisotropia geométrica, onde $\alpha 1$ e $\alpha 2$ identificam, respectivamente, os ângulos de maior e menor espalhamento do fenômeno em estudo, ϕ_1 e ϕ_2 são seus alcances, τ^2 é o efeito pepita e $\tau^2+\sigma^2$ o patamar.

Para modelarmos a anisotropia geométrica devemos proceder da seguinte forma:

(a) identificar as direções de maior ($\alpha 1$) e menor ($\alpha 2$) variabilidade espacial (lembre-se que $\alpha 2$ é ortogonal à $\alpha 1$);

(b) construir o semivariograma experimental para cada uma destas direções;

(c) ajustar os modelos teóricos de semivariograma utilizando uma função de correlação apropriada para expressar a estrutura de dependência espacial da variável em estudo para cada uma das direções acima citadas e identificar seus parâmetros estruturais (patamar, efeito pepita e alcance);

(d) uma vez definidos os modelos de semivariograma relativos às direções $\alpha 1$ e $\alpha 2$, $\gamma_{\alpha 1}(u)$ e $\gamma_{\alpha 2}(u)$ respectivamente, elaboramos um único modelo, $\gamma(u)$, para qualquer distância e direção de u, como especificado a seguir.

Sabemos que esta função de correlação depende de u, que é um vetor com componente distância e direção e que tem seu módulo decomposto da seguinte forma:

$$|u| = \sqrt{(u_{\alpha 1})^2 + (u_{\alpha 2})^2}\ , \qquad \textbf{(6)}$$

e ilustrado em **Figura 5**.

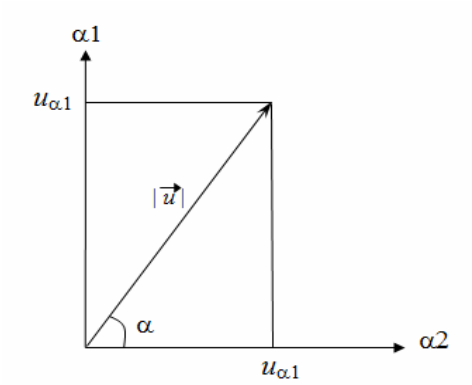


Figura 5. Decomposição genérica do vetor *u*.

Para a direção de análise $\alpha 1$, o vetor u não possui componente na direção $\alpha 2$, ou seja, para $\alpha 1$ temos que $\alpha = 90^o$ e, portanto, $u_{\alpha 1} = |u|.\text{sen}(90^o) = |u|$ e $u_{\alpha 2} = |u|.\cos(90^o) = 0$.

Normalizando **(6)** em relação ao alcance (ϕ), temos que:

$$\left|\frac{u}{\phi}\right| = \sqrt{\left(\frac{u_{\alpha 1}}{\phi}\right)^2 + \left(\frac{u_{\alpha 2}}{\phi}\right)^2} \qquad \textbf{(7)}$$

Como a componente $\left(\frac{u_{\alpha 2}}{\phi}\right)$ é sempre nula, podemos atribuir um alcance infinito na direção $\alpha 2$. Desta forma, a equação **(7)** pode ser reescrita na forma:

$$\left|\frac{u}{\phi}\right| = \lim_{a\to\infty} \sqrt{\left(\frac{u_{\alpha 1}}{\phi}\right)^2 + \left(\frac{u_{\alpha 2}}{a}\right)^2} \qquad \textbf{(8)}$$

Assim, os modelos normalizados dos semivariogramas relativos às direções $\alpha 1$ e $\alpha 2$ são definidos como:

$$\gamma_{\alpha 1}(u) = \lim_{a\to\infty} \tau^2 + \sigma^2 \left[1 - \rho \left(\sqrt{\left(\frac{u_{\alpha 1}}{\phi_1}\right)^2 + \left(\frac{u_{\alpha 2}}{a}\right)^2} \right) \right] \text{ e} \qquad \textbf{(9)}$$

$$\gamma_{\alpha 2}(u) = \lim_{a\to\infty} \tau^2 + \sigma^2 \left[1 - \rho \left(\sqrt{\left(\frac{u_{\alpha 1}}{a}\right)^2 + \left(\frac{u_{\alpha 2}}{\phi_2}\right)^2} \right) \right], \qquad \textbf{(10)}$$

em que $\alpha 1$ e $\alpha 2$ identificam, respectivamente, os ângulos de maior e menor espalhamento do fenômeno em estudo, ϕ_1 e ϕ_2 são seus alcances, τ^2 é o efeito pepita, $\tau^2+\sigma^2$ o patamar e $\lim_{a\to\infty} \rho \left(\sqrt{\left(\frac{u_{\alpha 1}}{\phi 1}\right)^2 + \left(\frac{u_{\alpha 2}}{a}\right)^2} \right)$ e $\lim_{a\to\infty} \rho \left(\sqrt{\left(\frac{u_{\alpha 1}}{a}\right)^2 + \left(\frac{u_{\alpha 2}}{\phi_2}\right)^2} \right)$ são as funções de correlação para cada direção.

Finalmente, uma vez definidos os modelos de semivariograma relativos às direções $\alpha 1$ e $\alpha 2$, elaboramos um único modelo para qualquer distância e direção de *u*, expresso através da seguinte equação:

$$\gamma(u) = \tau^2 + \sigma^2 \left[1 - \rho \left(\sqrt{\left(\frac{u_{\alpha 1}}{\phi_1}\right)^2 + \left(\frac{u_{\alpha 2}}{\phi_2}\right)^2} \right) \right], \qquad \textbf{(11)}$$

Um outro tipo de anisotropia que pode ser verificada é a denominada anisotropia zonal onde a amplitude permanece constante e o patamar varia de acordo com a direção. É um caso não muito freqüente nos fenômenos naturais. O mais comum é encontrarmos combinações de anisotropia geométrica e zonal, denominada anisotropia combinada ou mista, quando as várias direções resultam em diferentes semivariogramas, variando tanto a amplitude quanto o patamar (**Figura 6**).

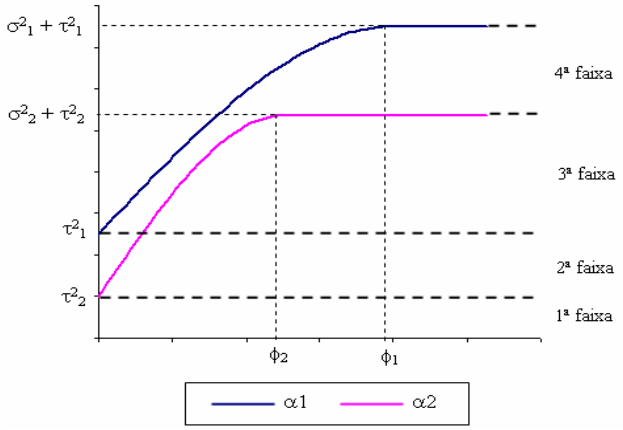


Figura 6. Anisotropia combinada e sua decomposição (em faixas) para modelagem, onde $\alpha 1$ é o ângulo de anisotropia, ϕ_1 seu alcance, τ^2_1 seu efeito pepita e $\tau^2_1+\sigma^2_1$ seu patamar, $\alpha 2$ é ângulo de menor espalhamento do fenômeno, ϕ_2 seu alcance, τ^2_2 seu efeito pepita e $\tau^2_2+\sigma^2_2$ seu patamar.

Para modelarmos a anisotropia combinada seguimos inicialmente os mesmos procedimentos (a), (b) e (c) citados na modelagem da anisotropia geométrica, mas, para gerar o modelo completo utilizamos uma técnica que consiste em dividirmos o gráfico do semivariograma teórico em faixas, conforme ilustrado em **Figura 6**, de modo que cada faixa represente somente uma anisotropia geométrica. Desta forma, teremos quatro estruturas:

$$\gamma_1(u) = \tau_2^2 \qquad \textbf{(12)}$$

$$\gamma_2(u) = \lim_{\varepsilon \to 0}(\tau_1^2 - \tau_2^2)\left[1 - \rho\left(\sqrt{\left(\frac{u_{\alpha 1}}{\varepsilon}\right)^2 + \left(\frac{u_{\alpha 2}}{\phi_2}\right)^2}\right)\right] \qquad \textbf{(13)}$$

$$\gamma_3(u) = (\sigma_2^2 + \tau_2^2 - \tau_1^2)\left[1 - \rho\left(\sqrt{\left(\frac{u_{\alpha 1}}{\phi_1}\right)^2 + \left(\frac{u_{\alpha 2}}{\phi_2}\right)^2}\right)\right] \qquad \textbf{(14)}$$

$$\gamma_4(u) = \lim_{a \to \infty}(\sigma_1^2 + \tau_1^2 - \sigma_2^2 + \tau_2^2)\left[1 - \rho\left(\sqrt{\left(\frac{u_{\alpha 1}}{\phi_1}\right)^2 + \left(\frac{u_{\alpha 2}}{a}\right)^2}\right)\right], \qquad \textbf{(15)}$$

O modelo completo e consistente para qualquer distância e direção do vetor *u* resume-se na soma das quatro estruturas acima definidas: $\gamma(u) = \gamma_1(u) + \gamma_2(u) + \gamma_3(u) + \gamma_4(u)$.

Os métodos de ajuste dos modelos podem ser divididos em dois grandes grupos:

A) <u>ajuste dos modelos ao semivariograma experimental</u> – os métodos de ajuste deste grupo são: método dos Mínimos Quadrados Ordinários (Ordinary Least Squares - OLS), método dos Mínimos Quadrados Ponderados (Weight Least Squares – WLS) e método de ajuste denominado de “a sentimento”;

B) <u>método de ajuste de um modelo direto aos dados</u> – Método da Máxima Verossimilhança (Maximum Likelihood – ML).

Explicitaremos melhor somente os métodos dos Mínimos Quadrados Ordinários e o dos Mínimos Quadrados Ponderados seguindo os procedimentos apresentados por JIAN et al. (1996).

Seja $\hat{\gamma}(u)$ a forma vetorial de um semivariograma experimental contendo *k* estimativas para valores incrementados do *lag*, ou seja, $\hat{\gamma}(u) = [\hat{\gamma}(u_1), \hat{\gamma}(u_2), ..., \hat{\gamma}(u_k)]$ e seja $\gamma(u, \theta) = [\gamma(u_1, \theta), \gamma(u_2, \theta), ..., \gamma(u_k, \theta)]$ um vetor com valores do modelo de semivariograma de interesse com parâmetros desconhecidos $\theta = [\theta_1, \theta_2, ..., \theta_p]$.

Seja, também, *R*, um valor que representa a diferença da soma de quadrados entre os valores observados e os estimados pelo modelo, expresso por:

$$R = \left[\hat{\gamma}(u) - \gamma(u,\theta)\right]^T V^{-1} \left[\hat{\gamma}(u) - \gamma(u,\theta)\right]$$

O melhor conjunto de parâmetros é aquele que minimiza *R*.

Na estimativa dos parâmetros do semivariograma via Mínimos Quadrados Ordinários, *V* é igual a matriz identidade e, portanto, a minimização de *R* é direta. Outra particularidade é que este método supõe que as diferenças são independentes, normalmente distribuídas e que todos os valores estimados têm a mesma variância.

A estimativa pelo Método dos Mínimos Quadrados Ponderados considera para *V* somente os termos da diagonal principal da matriz de variância/covariância. Este método utiliza iteração, que é rápida, pois todos os elementos fora da diagonal principal são assumidos como sendo zero, logo a inversão desta matriz é trivial (a inversão de uma matriz diagonal é também uma matriz diagonal). Neste caso, cada diferença é ponderada diretamente pelo inverso da variância do semivariograma experimental.

O método de ajuste "a sentimento" é subjetivo e depende da experiência do pesquisador, já que consiste em um ajuste visual do modelo selecionado aos pontos do semivariograma experimental.

Finalmente, o método de ajuste direto aos dados, Método da Máxima Verossimilhança, quando aplicado a amostras grandes, nos fornece estimadores não viciados e eficientes. A idéia neste caso é obter, a partir de uma amostra, o estimador "mais verossímel" dos parâmetros de um certo modelo probabilístico.

A avaliação do desempenho de cada modelo geoestatístico pode ser efetuada através do critério de informação de Akaike, AIC. A estimava deste critério é expressa por:

$$\hat{AIC} = n \ \ln\left(\frac{R_m}{n}\right) + 2p\,,$$

onde *n* é o número de pontos do semivariograma experimental e *p* é o número de parâmetros do modelo.

Decidiremos, então, a escolha dos parâmetros do modelo com o menor valor de AIC, o modelo mais parcimonioso.

Finalmente, após o ajuste do modelo proposto, procedemos a análise geoestatística executando a validação do modelo, uma técnica que nos permite avaliar a adequação do modelo escolhido, ou seja, nos permite avaliar o grau de incerteza sobre os parâmetros utilizados.

Neste procedimento cada valor original é removido do domínio espacial e, usando-se os demais, um novo valor é estimado para este ponto através dos parâmetros ajustados ao modelo do semivariograma.

A validação não nos permite provar que o modelo adotado é o mais correto, mas sim que ele não é inteiramente incorreto.

3 - MATERIAL E MÉTODOS

3.1 Material

O Banco de Dados

Os dados a serem aqui analisados são provenientes de um projeto intitulado "*Estudo da relação entre indicadores entomológicos para Aedes (Stegomyia) aegypti obtidos de armadilhas adulticidas, de oviposição e de coleta de adultos, em área da região noroeste do estado de São Paulo*" (CHIARAVALLOTI-NETO et al. (2004), Faculdade de Medicina de São José do Rio Preto) que possui como objetivo geral avaliar, para o *Aedes aegypti*, as relações entre os indicadores entomológicos obtidos através de armadilhas adulticidas e de oviposição e através de coletas de adultos e os indicadores climáticos na cidade de Mirassol, lotada na região noroeste do estado de São Paulo.

Nesta dissertação vamos trabalhar com os dados das armadilhas adulticidas, M*osquitrap*, buscando encontrar modelos que expliquem bem os dados observados.

A *Mosquitrap* é uma armadilha que captura a fêmea grávida do pernilongo. Foi desenvolvida pelo Professor Álvaro Eiras do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais (EIRAS, 2002). Em seu interior, ela possui uma substância viscosa que impede a fêmea grávida de sair. Esta substância, que contém o feromônio para atrair o inseto, também foi desenvolvida por EIRAS e é denominada Atr*Aedes*.

As armadilhas *Mosquitrap* foram distribuídas em duas áreas de estudo com amostragens distintas. A primeira área de análise, aqui denominada **A1**, é constituída por 100 quadras e a segunda área, intitulada **A4**, engloba 30 quadras. Em cada quarteirão da área **A1** foram instaladas armadilhas adulticidas com alternância da direção, ora na direção norte-sul, ora na direção leste-oeste. Assim, nesta área **A1**, temos uma amostragem de 100 observações georeferenciadas, uma *Mosquitrap* por quadra. Na segunda área de análise, **A4**, temos uma amostragem de 120 observações georeferenciadas, 4 armadilhas adulticidas por quadra. Este cenário está ilustrado em **Figura 7**.

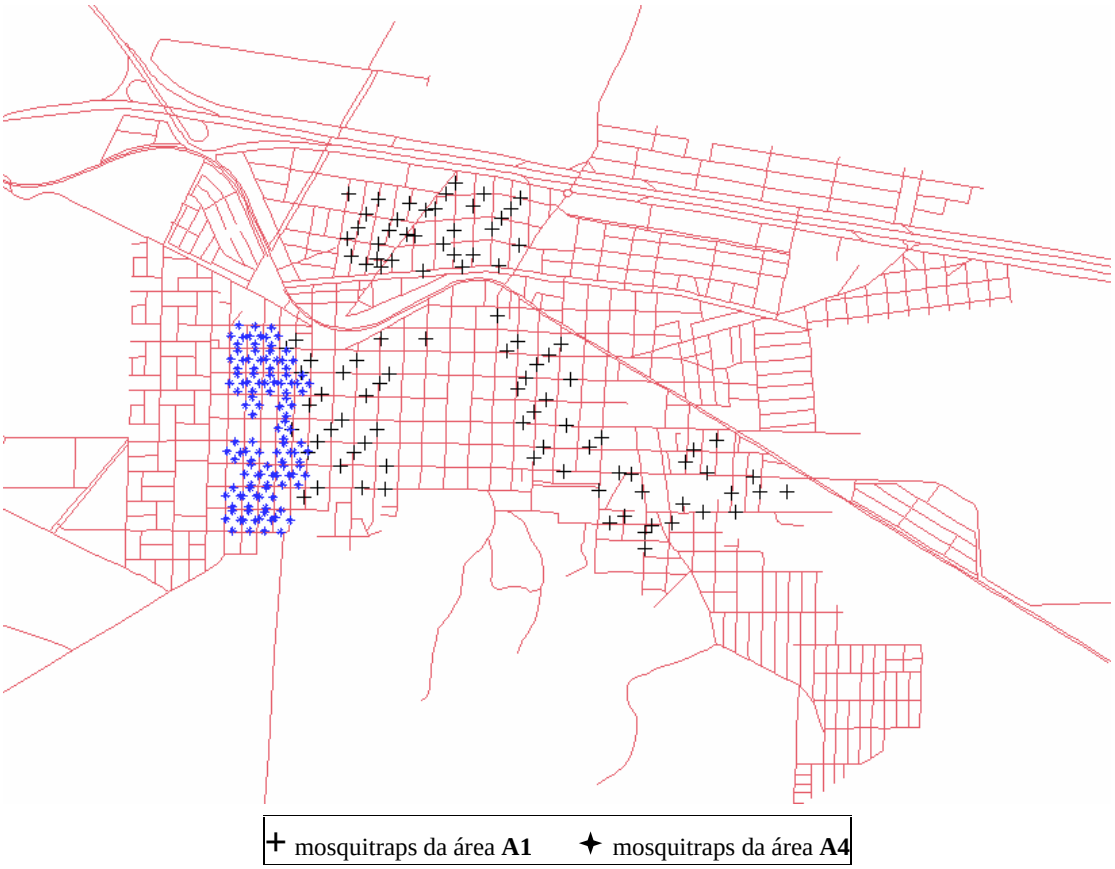


Figura 7. Mapa da cidade de Mirassol/SP com a localização das armadilhas para cada área de análise.

Para tentar explicar a presença/comportamento das fêmeas *Aedes* capturadas foram ainda coletados dados das seguintes variáveis meteorológicas: temperatura mínima, temperatura máxima, temperatura média e pluviosidade.

Sumariamente, nesta análise trabalhamos com dados das armadilhas adulticidas de casas situadas em 100 quarteirões, uma armadilha por quadra - área **A1,** e 30 quarteirões, quatro armadilhas por quadra - área **A4**, coletados durante 23 semanas (**Tabela 2**). Neste contexto, para a área **A1** temos um total de 2.300 observações sendo a quadra a unidade amostral e, na área **A4**, temos 2.760 valores observados e o domicílio considerado como unidade amostral.

Assim, a variável de análise é o número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas e as covariáveis são temperatura mínima, temperatura máxima, temperatura média e pluviosidade.

Vale salientar, ainda, que possuímos informações das variáveis meteorológicas referentes ao dia da coleta e de até 21 dias atrás, portanto, temos 88 variáveis explicativas que podem ser incorporadas ao modelo.

Tabela 2. Datas das coletas para cada semana e área de análise

Semana	Coleta	
	A1	**A4**
1	22/11 a 25/11/04	22/11/04
2	29/11 a 02/12/04	29/11/04
3	06/12 a 09/12/04	06/12/04
4	13/12 a 16/12/04	13/12/04
5	20/12 a 23/12/04	20/12/04
6	27/12 a 30/12/04	27/12/04
7	03/01 a 06/01/05	03/01/05
8	10/01 a 13/01/05	10/01/05
9	17/01 a 20/01/05	17/01/05
10	24/01 a 27/01/05	24/01/05
11	31/01 a 03/02/05	31/01/05
12	07/02 a 10/02/05	07/02/05
13	14/02 a 17/02/05	14/02/05
14	21/02 a 24/02/05	21/02/05
15	28/02 a 03/03/05	28/02/05
16	07/03 a 10/03/05	07/03/05
17	14/03 a 17/03/05	14/03/05
18	21/03 a 24/03/05	21/03/05
19	28/03 a 31/03/05	28/03/05
20	04/04 a 07/04/05	04/04/05
21	11/04 a 14/04/05	11/04/05
22	18/04 a 21/04/05	18/04/05
23	25/04 a 28/04/05	25/04/05

3.2 Metodologia

Para relacionarmos a quantidade de *Aedes* capturadas com as variáveis meteorológicas, por se tratar de dados de contagem, inicialmente podemos supor que a variável resposta assume uma distribuição de Poisson e a função de ligação logarítmica.

Em uma primeira análise, para ambas as áreas de amostragem, **A1** e **A4**, ajustamos modelos utilizando a metodologia de Modelos Lineares Generalizados e, posteriormente, somente para a área **A1**, incorporamos a correlação temporal através da modelagem via teoria de Equações de Estimação Generalizadas, uma vez que identificamos uma estrutura de correlação auto-regressiva de ordem 1. Na área **A4**, possivelmente devido ao esquema de amostragem utilizado, não detectamos uma estrutura de correlação temporal e, portanto, não efetuamos ajustes via EEG para explicarmos a presença do *Aedes* nesta área.

Somente para facilitar a apresentação dos resultados denominamos estes modelos por:

- **Mosquitrap_A1 – MLG**;
- **Mosquitrap_A1 – EEG**;
- **Mosquitrap_A4 – MLG**.

Através desta nomenclatura estamos especificando, para cada um deles, a variável resposta objeto da análise (**Y** = mosquitrap, ou seja, o número de fêmeas *Aedes aegypti* capturadas pela armadilha adulticida mosquitrap), a área de amostragem em questão (**A1** ou **A4**) e a teoria utilizada para o ajuste do modelo (MLG ou EEG).

As seleções de variáveis explicativas dos modelos acima citados foram efetuadas através das funções stepwise_glm e stepwise_gee desenvolvidas em *R*. Elaboramos estas funções, pois não encontramos procedimentos stepwise desenvolvidos para modelagem por MLG com termos fixos, como no nosso caso, onde a quadra foi considerada como bloco para a área **A1** e o domicílio para a área **A4**. Stepwise_gee foi desenvolvida, pois, até então, não tínhamos conhecimento da implantação deste método em *R* ou em SAS, softwares disponíveis no DEs/UFSCar.

Utilizamos o procedimento stepwise uma vez que seu uso é indicado quando o número de covariáveis é grande, e adotamos como critério de seleção o valor *p*, o nível descritivo do teste, tendo sido determinado um valor *p* de 5% para a entrada e 10% para a retirada da covariável.

Para o uso da função stepwise_glm$^{(\text{Apêndice C – C1})}$ é necessário fornecermos os seguintes parâmetros:

- **Y**, a variável resposta;
- VE, um vetor contendo o nome (**tipo texto**, portanto, especificado entre aspas duplas) de todas as variáveis explicativas que podem ser introduzidas no modelo;
- PDADOS, o nome do objeto que contém o banco de dados;
- FAMILIA, identificando a distribuição da variável resposta e a função de ligação utilizada. Por exemplo, FAMILIA = poisson(link=log);
- NIVEL, o nível de significância utilizado para introduzir a variável no modelo;
- NIVEL_RET, o nível de significância utilizado para retirar a variável do modelo;
- TERMO_OBR, um vetor contendo o nome (**tipo texto**, portanto, especificado entre aspas duplas) de termos obrigatórios do modelo. Por exemplo, TERMO_OBR = c(“factor(quadra)”) ou TERMO_OBR = c(“X1”, “X2”), se não houver covariáveis fixas, devemos definir um valor inicial para este parâmetro da seguinte forma: TERMO_OBR = c();

- NUM_PAR, o número de parâmetros fixos do modelo. Por exemplo, se for somente o intercepto, NUM_PAR = 1. Se tiver um termo obrigatório do tipo fator, lembre-se que o *R* incorpora um dos níveis do fator como intercepto, assim, como no nosso caso em que consideramos a quadra como fator (100 quadras), temos NUM_PAR = 100 já que uma quadra é o intercepto. Se, TERMO_OBR = c("X1", "X2"), então NUM_PAR = 3, o intercepto e duas covariáveis, X1 e X2, fixas no modelo.

Para o uso da função stepwise_gee(Apêndice C – C2) são solicitados os seguintes parâmetros:

- **Y**, a variável resposta;
- VE, um vetor contendo o nome (**tipo texto**, portanto, especificado entre aspas duplas) de todas as variáveis explicativas que podem ser introduzidas no modelo;
- PID, a unidade de repetição;
- PDADOS, o nome do objeto que contém o banco de dados;
- FAMILIA, identificando a distribuição da variável resposta e a função de ligação utilizada (por exemplo, FAMILIA = poisson(link="log"));
- ESTRUTURA, identificando a estrutura da matriz de correlação de trabalho (**do tipo texto**). Por exemplo, ESTRUTURA = "ar1";
- NIVEL, o nível de significância utilizado para introduzir a variável no modelo;
- NIVEL_RET, o nível de significância utilizado para retirar a variável do modelo;
- TERMO_OBR, um vetor contendo o nome (**tipo texto**, portanto, especificado entre aspas duplas) de termos obrigatórios do modelo;
- NUM_PAR, o número de parâmetros fixos do modelo.

A função stepwise_glm desenvolvida ajusta os modelos através da função *glm* do *R* e, portanto, as famílias e funções de ligações se restringem àquelas presentes na função *glm*.

A função stepwise_gee utiliza no ajuste a função *geeglm*, presente no pacote geepack do *R* e, assim, as famílias, as funções de ligações e as estruturas da matriz de correlação de trabalho se restringem àquelas presentes na função *geeglm*.

Como uma tentativa de identificarmos o raio de correlação das amostras e a direção privilegiada da variabilidade espacial do número de fêmeas *Aedes* capturadas por armadilhas adulticidas, ou melhor, para identificarmos sua estrutura de correlação espacial, prosseguimos

nosso estudo efetuando modelagens via teoria de Geoestatística para que ações de controle mais precisas possam ser realizadas.

As modelagens via teoria de Geoestatística foram efetuadas através do software *Spring*, Sistema para Processamento de Informações Georeferenciadas, um banco de dados geográfico desenvolvido pelo INPE (Instituto Nacional de Pesquisas Espaciais). Com esta análise obtivemos os alcances de maior e menor espalhamento do fenômeno em estudo, denominados respectivamente de **Alcance1** e **Alcance2**.

Em uma última etapa deste estudo, ajustamos modelos para tentar relacionar os alcances de correlação das amostras com as variáveis meteorológicas. Para esta nova variável resposta, do tipo contínua, assumimos uma distribuição Gama e uma função de ligação logarítmica. Para os alcances da área **A1**, uma vez que não identificamos a existência de uma correlação temporal, ajustamos modelos via teoria de MLG e utilizamos a função stepwise_glm para selecionarmos as covariáveis, com níveis de entrada e retirada iguais a 0,5%. Os alcances da área **A4** também não nos revelaram a existência de uma estrutura de dependência temporal e, por isso, ajustamos modelos via teoria de MLG utilizando a função stepwise_glm para identificarmos as variáveis explicativas mais significativas, também com níveis de entrada e retirada iguais a 0,5%. Ao tentarmos identificar as covariáveis que pudessem explicar os alcances da área **A4,** via função *glm* do R, verificamos que o algoritmo não convergia e, portanto, não conseguimos, neste momento, identificar um modelo para justificar os resultados obtidos para as variáveis respostas em questão (alcance1 e alcance2 – área **A4**).

Assim como anteriormente, para facilitar a apresentação dos resultados, vamos denominar estes últimos modelos analisados por:

- **Alcance1_A1 – MLG**;

- **Alcance2_A1 – MLG**.

Novamente esta nomenclatura foi utilizada para que se identifique, para cada um deles, a variável resposta objeto da análise (**Alcance1** ou **Alcance2**, representando, respectivamente, os alcances de maior e menor espalhamento do fenômeno em estudo), a área de amostragem (**A1**) e a teoria utilizada para o ajuste do modelo (MLG).

Os modelos considerados neste estudo são apresentados a seguir.

3.2.1 Aplicação dos Modelos Lineares Generalizados na modelagem do número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas, modelos: Mosquitrap_A1 – MLG e Mosquitrap_A4 - MLG

Seja **Y** o número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas. Assumimos que variável resposta possui uma distribuição de Poisson (**Y** ~ Poisson(μ)), com função densidade probabilidade (f.d.p.) dada por:

$$f(\mathbf{y};\mu) = \frac{\mu^{\mathbf{y}} e^{-\mu}}{\mathbf{y}!} I_A(\mathbf{y}), \quad \mu > 0; \quad A = \{1,2,...\}.$$

Temos, então,

$$f(\mathbf{y};\mu) = \exp\{\mathbf{y}\ln(\mu) - \mu - \ln(\mathbf{y}!)\} I_A(\mathbf{y}),$$

obtendo-se

$$a(\phi) = 1,\ \theta = \ln \mu \Leftrightarrow e^{\theta} = \mu,\ b(\theta) = e^{\theta} \text{ e } c(\mathbf{y};\theta) = -\ln \mathbf{y}!.$$

Desta forma, mostramos que a distribuição de Poisson pertence à família exponencial na forma dada por equação **(1)**.

Assim, temos que a esperança de **Y** é dada por

$$E[\mathbf{Y}] = b'(\theta) = e^{\theta} = \mu,$$

com variância expressa por

$$Var[\mathbf{Y}] = a(\phi)b''(\theta) = a(\phi)V(\mu) = e^{\theta} = \mu$$

onde $V(\mu) = e^{\theta} = \mu$ é sua função de variância.

Como assumimos uma função de ligação logarítmica, ou seja, $\eta = g(\mu) = \ln(\mu)$, temos que o modelo log-linear, considerando-se a unidade amostral como um fator (sendo a quadra para a área **A1** e o domicílio para a área **A4**), será expresso por:

$$\mu = \exp\{\text{fator(unidade\_amostral)}+\eta\} = \exp\{\text{fator(quadra)}+\mathbf{X}\beta\},$$

onde

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n)'$ é a matriz do modelo;

$\beta = (\beta_1, \beta_2, \beta_3,..., \beta_p)'$ é o vetor de parâmetros desconhecidos, onde, no nosso caso, podemos ter até 89 parâmetros;

$\eta = (\eta_1, \eta_2,..., \eta_n)'$ é o preditor linear;

$\mu = (\mu_1, \mu_2, \mu_3,..., \mu_n)'$, é o vetor de médias, ou seja, $E(Y_i) = \mu_i$, $i = 1,..., n$, com $n = 2.300$ observações para a área **A1** (100 quadras analisadas durante 23 semanas) e $n = 2.760$ observações para a área **A4** (120 domicílios analisados durante 23 semanas).

A função deviance para o modelo de Poisson é expressa por:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left\{y_i\left(\hat{\theta}_i^0 - \hat{\theta}_i\right) + \left(b(\hat{\theta}_i) - b(\hat{\theta}_i^0)\right)\right\}$$

$$= 2\sum_{i=1}^{n} w_i\left(y_i\,[log\, y_i - ln\,\hat{\mu}_i\,] - y_i + \hat{\mu}_i\right) =$$

$$= 2\sum_{i=1}^{n} w_i\left(y_i\, log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right)$$

A estatística de Pearson X^2 generalizada é dada por:

$$X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \sum_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}\,.$$

3.2.2 Aplicação das Equações de Estimação Generalizadas na modelagem do número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas, modelo: Mosquitrap_A1 – EEG

Seja a variável aleatória **Y** o número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas. Com a aplicação desta teoria consideramos as variáveis respostas, Y_{ij}, i = 1, 2,..., K, j = 1, 2,..., n_i, representando a j-ésima semana medida da i-ésima unidade amostral (para a área **A1** temos k = 100 quadras e n_i = 23), como sendo dependentes.

A identificação da estrutura de correlação foi efetuada através do correlograma, uma representação gráfica das autocorrelações de um conjunto de dados, que nos revelou a existência de uma estrutura auto-regressiva de ordem 1, AR(1), para a área **A1**.

Assim, temos que a matriz de trabalho $\mathbf{R}_i(\alpha)$ apresenta a seguinte estrutura de correlação entre as medidas repetidas:

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots & \alpha^{22} \\ \alpha & 1 & \alpha & \alpha^2 & \cdots & \alpha^{21} \\ \alpha^2 & \alpha & 1 & \alpha & \cdots & \alpha^{20} \\ \vdots & \vdots & \vdots & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ \alpha^{22} & \alpha^{21} & \alpha^{20} & \cdots & \cdots & 1 \end{pmatrix}_{23x23},$$

onde i = 1, 2, ..., 100 quadras e $\hat{\alpha} = \frac{1}{k\phi}\sum_{i=1}^{K}\frac{1}{n_i - 1}\sum_{j\le n_i - 1} r_{ij} r_{i,j+1}$, com $r_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{V(\hat{\mu}_{ij})}}$ (resíduo de Pearson).

Como k = 100 quadras, n_i = 23 semanas, ϕ = 1 e $V(\hat{\mu}_{ij}) = \hat{\mu}_{ij}$ a função de variância para a distribuição de Poisson, temos que:

$$\hat{\alpha} = \frac{1}{100}\sum_{i=1}^{100}\frac{1}{23-1}\sum_{j\le 22} r_{ij} r_{i,j+1} = \frac{1}{2200}\sum_{i=1}^{100}\sum_{j\le 22} r_{ij} r_{i,j+1}\,,$$

onde $r_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$.

3.2.3 Aplicação da Geoestatística para identificar a estrutura de dependência espacial do número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas

Para cada uma das 23 semanas de cada área de análise (**A1** e **A4**) detectamos inicialmente a presença ou não da anisotropia através do mapa de semivariograma, ou semivariograma de superfície.

Após identificarmos as direções de maior e menor variabilidade espacial, prosseguimos a análise construindo o semivariograma experimental para cada uma destas direções. Neste estudo temos uma amostragem irregularmente distribuída nas duas áreas de análise e, desta forma, para o cálculo do semivariograma experimental definimos os **parâmetros do *lag***, a distância de tolerância Δu para o espaçamento u entre os pares de amostras, e os **parâmetros de direção**, um ângulo de tolerância $\Delta\alpha$ para a direção α considerada.

Nesta etapa, alternadamente, construímos o semivariograma experimental e ajustamos diversos modelos, da seguinte forma: alterávamos os **parâmetros do *lag*** (nº do *lag*, incremento e tolerância) e construíamos o semivariograma experimental, em seguida, ajustávamos um modelo especificando uma determinada função de correlação e analisávamos o resultado produzido pelo *Spring* em uma tela denominada "**Modelo de Ajuste**", ilustrada em **Figura 8**. Esta tela apresentava graficamente o ajuste do modelo teórico escolhido e os *lags* do semivariograma experimental, nos permitindo avaliar visualmente se o ajuste é ou não satisfatório. Além destes resultados, o *Spring* também apresenta um conjunto de informações na tela "**Relatório de Dados**", tais como, função de correlação escolhida e os parâmetros estruturais do modelo: efeito pepita, contribuição e alcance, além do valor de Akaike, indicador do ajuste realizado. Neste ponto, selecionávamos os valores dos parâmetros estruturais com menor valor de Akaike para compará-lo com outros ajustes realizados. Escolhido os valores dos parâmetros estruturais do semivariograma, calculávamos, então, o grau de aleatoriedade presente nos dados (E = patamar/efeito pepita) para verificarmos se possuíamos um modelo de pepita pura, situação na qual a aplicação da análise semivariográfica não se aplica.

Prosseguimos, então, seguindo os procedimentos acima citados e, após identificarmos o modelo com a função de correlação suposta adequada para expressar a estrutura de dependência espacial do fenômeno em estudo e os parâmetros estruturais do semivariograma teórico, definimos e validamos os modelos de semivariograma para cada uma das semanas de

análise. Vale salientar que o modelo escolhido foi aquele que apresentou o menor valor de Akaike, já que este valor nos indica que o modelo de ajuste é mais preciso.

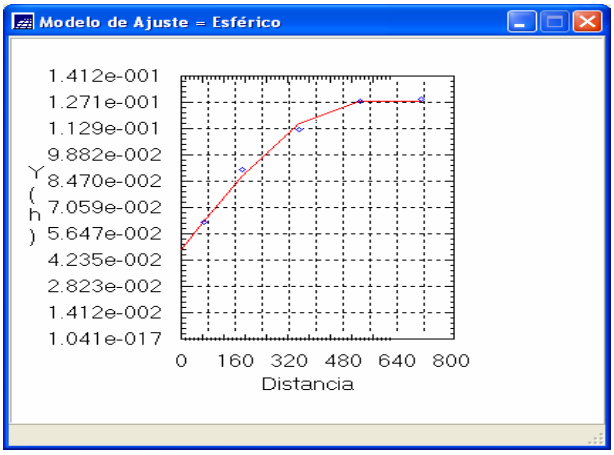


Figura 8. Tela do Spring denominada "Modelo de Ajuste" que apresenta graficamente o ajuste do modelo teórico escolhido (linha) e os *lags* do semivariograma experimental (pontos)

É importante ressaltar que a modelagem ou ajuste do semivariograma foi realizada no *Spring* no modo automático, que utiliza o algoritmo de JIAN et al. (1996), baseado no método dos mínimos quadrados para estimar os parâmetros estruturais do modelo.

A validação do modelo no *Spring*, realizada para avaliar grau de incerteza sobre os parâmetros utilizados, ou seja, avaliar o erro da estimativa, resume-se em re-estimar os valores conhecidos através dos parâmetros ajustados ao modelo do semivariograma. Para que conclusões possam ser definidas, este módulo nos fornece o diagrama espacial do erro, histograma do erro, estatísticas do erro, o diagrama dos valores observados x estimados e resultados numéricos.

3.2.4 Aplicação dos Modelos Lineares Generalizados na modelagem dos alcances de maior e menor espalhamento do número de fêmeas *Aedes aegypti* capturadas em armadilhas adulticidas na área de amostragem A1, modelos: Alcance1_A1 – MLG e Alcance2_A1 - MLG

Assumimos independência das observações uma vez que os correlogramas construídos para ambos alcances não nos indicou a existência de dependência temporal.

Sejam **X** e **Z** as variáveis aleatórias dos alcances **Alcance1** e **Alcance2**, representando, respectivamente, o raio de maior e menor espalhamento do fenômeno em estudo. A distribuição padrão a ser assumida é a Gama, assim, $\mathbf{X} \sim G_x(\mu_x,\nu_x)$ e $\mathbf{Z} \sim G_z(\mu_z,\nu_z)$, com função densidade probabilidade (f.d.p.) dada por:

$$f(\mathbf{x};\mu_x,\nu_x) = \frac{\left(\frac{\nu_x}{\mu_x}\right)^{\nu_x}}{\Gamma(\nu_x)} x^{\nu_x-1} \exp\left(-\frac{x\nu_x}{\mu_x}\right) I_A(\mathbf{x}), \quad \mu_x > 0, \quad \nu_x > 0; \quad A = \Re^+,$$

onde $\Gamma(\nu_x) = \int_0^\infty u^{\nu_x - 1} e^{-u} du, \ \ \nu_x > 0$.

Temos, então,

$$f(\mathbf{x}; \mu_x, \nu_x) = \exp\left\{ \nu_x \left(-\frac{x}{\mu_x} - \ln \mu_x \right) + \nu_x \ln(x\nu_x) - \ln x - \ln \Gamma(\nu_x) \right\} I_A(\mathbf{x})$$

obtendo-se

$a(\phi) = \frac{1}{\nu_x}$, $\theta = -\frac{1}{\mu} \Leftrightarrow -\frac{1}{\theta} = \mu_x$, b(θ) = - ln(-θ) e c(x; θ) = ν_x ln(xν_x) – ln x – ln Γ(ν_x).

Mostramos, portanto, que a distribuição Gama pertence à família exponencial na forma dada por equação **(1)**.

Desta forma, temos que a esperança de **X** é dada por:

$$E[\mathbf{X}] = b'(\theta) = -\frac{1}{\theta} = \mu_x,$$

com variância expressa por

$$Var[\mathbf{X}] = a(\phi)b''(\theta) = a(\phi)V(\mu_x) = \frac{\mu_x^{\,2}}{\nu_x},$$

onde $V(\mu_x) = \mu_x^2$ é sua função de variância.

Como assumimos uma função de ligação logarítmica, ou seja, $\eta = g(\mu) = \ln(\mu)$, temos que o modelo log-linear, considerando-se a quadra como um fator, será expresso por:

$$\mu = \exp\{\text{fator(quadra)}+\eta\} = \exp\{\text{fator(quadra)}+\mathbf{C}\beta\},$$

onde

C = (**c**$_1$, **c**$_2$,...,**c**$_n$)' é a matriz do modelo;

β = (β_1, β_2, β_3,..., β_p)' é o vetor de parâmetros desconhecidos, onde, no nosso caso, podemos ter até 89 parâmetros;

η = (η_1, η_2,..., η_n)' é o preditor linear;

μ = (μ_{x1}, μ_{x2}, μ_{x3},..., μ_{xn})', é o vetor de médias, ou seja, E(X_i)= μ_i, i = 1,..., n, com n = 1600 observações, já que temos 100 quadras analisadas durante 16 semanas – 7 semanas não puderam ser modeladas via teoria da Geoestatística.

A função deviance para o modelo Gama é expressa por:

$$D(\mathbf{x};\hat{\boldsymbol{\mu}}) = 2v\sum_{i=1}^{n}\left\{x_i\left(\hat{\theta}_i^0 - \hat{\theta}_i\right) + \left(b(\hat{\theta}_i) - b(\hat{\theta}_i^0)\right)\right\}$$

$$= 2v_x\sum_{i=1}^{n} w_i\left(x_i\left[-\frac{1}{x_i} - \left(-\frac{1}{\hat{\mu}_i}\right)\right] - \left(-\ln\left(\frac{1}{x_i}\right)\right) + \left(-\ln\left(\frac{1}{\hat{\mu}_i}\right)\right)\right) =$$

$$= 2v_x\sum_{i=1}^{n} w_i\left(-\ln\left(\frac{x_i}{\hat{\mu}_i}\right) + \frac{x_i - \hat{\mu}_i}{\hat{\mu}_i}\right)$$

A estatística de Pearson X^2 generalizada é dada por:

$$X_x^2 = \sum_{i=1}^{n} w_i \frac{(x_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \sum_{i=1}^{n} w_i \frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}.$$

Por simetria, temos E[**Z**] = b'(θ) = $-\frac{1}{\theta} = \mu_z$ e Var[**Z**] = a(ϕ)b''(θ) = a(ϕ)V(μ_z) = $\frac{\mu_z^{\,2}}{v_z}$.

4 – RESULTADOS E DISCUSSÃO

4.1 Resultados dos modelos elaborados para estudar a variável resposta Y, o número de fêmeas *Aedes* capturadas em armadilhas adulticidas – áreas A1 e A4

Em uma análise descritiva inicial, construímos os histogramas da variável resposta **Y** das áreas **A1** e **A4**, e, em seguida, comparamos os valores observados destas amostras com uma distribuição de Poisson, através da sobreposição dos histogramas, conforme ilustrado em **Figura 9** e **Figura 10**.

Através dos resultados gráficos obtidos, concluímos que a suposição da distribuição de Poisson para a variável dependente **Y**, pode ser considerada nesta análise inicial, havendo, ainda, a necessidade de confirmar a adequação da função de ligação logarítmica.

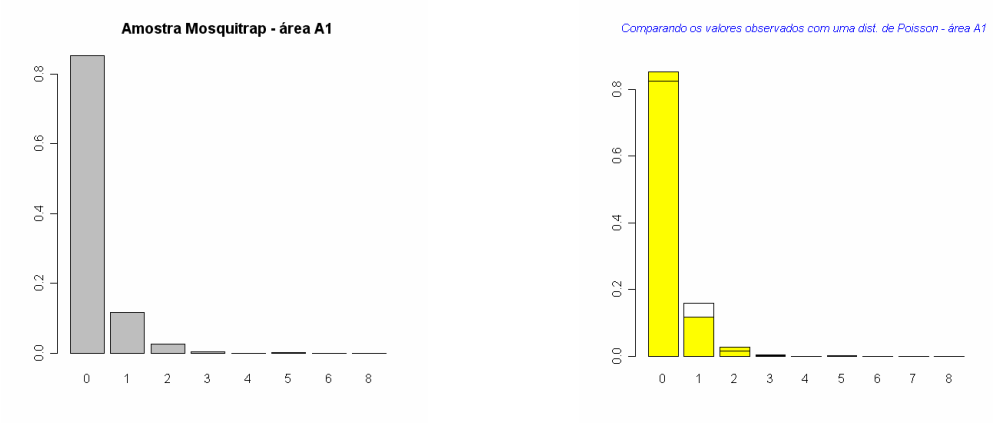


Figura 9. Histograma da variável resposta Y – área A1 – à esquerda e sobreposição de histogramas: amostra observada (transparente) e distribuição de Poisson (cinza) – à direita.

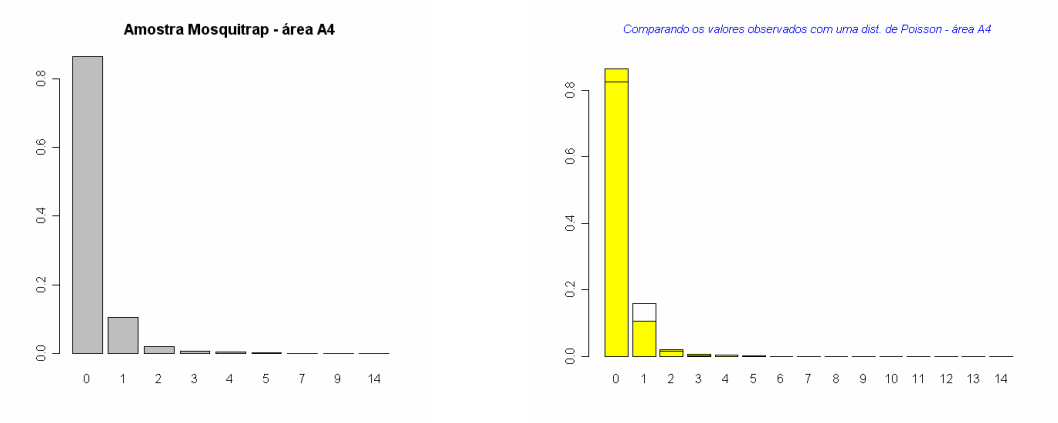


Figura 10. Histograma da amostra mosquitrap – área A4 – à esquerda e sobreposição de histogramas: amostra observada (transparente) e distribuição de Poisson (cinza) – à direita.

Em seguida, para tentar relacionar **Y**, a quantidade de fêmeas *Aedes* capturadas em armadilhas adulticidas, com as variáveis meteorológicas, ajustamos, então, os modelos: **Mosquitrap_A1 – MLG**, onde a quadra foi considerada bloco e **Mosquitrap_A4 - MLG**, onde

os domicílios foram considerados como blocos; e o modelo **Mosquitrap_A1 – EEG,** sendo a quadra a unidade de repetição.

4.1.1 Ajuste do modelo Mosquitrap_A1 - MLG

O modelo obtido, as estimativas dos parâmetros e seus respectivos desvios padrão foram:

$$\hat{\eta} = \underset{(0,93)}{0,59} - \underset{(0,7)}{0,69}Q_{101} + ... + \underset{(0,46)}{1,15}Q^{*}_{125} - \underset{(1,08)}{1,79}Q^{\bullet}_{126} + ... + \underset{(0,48)}{0,98}Q^{*}_{41} + ... + \underset{(0,49)}{0,87}Q^{\bullet}_{79} - \underset{(1,08)}{1,79}Q^{\bullet}_{80} - ... - \underset{(1,08)}{1,79}Q^{\bullet}_{94} - ... +$$
$$+ \underset{(0,002)}{0,01}pluv^{***}_{18} - \underset{(0,01)}{0,05}t\max^{***}_{9} + \underset{(0,005)}{0,01}pluv^{\bullet}_{13} - \underset{(0,01)}{0,05}t\max^{***}_{5} + \underset{(0,03)}{0,08}t\min^{**}_{4}$$

onde nível: *** 0,001; ** 0,01; * 0,05; • 0,1

Verificamos, também, a adequabilidade da função de ligação adicionando $\hat{\boldsymbol{\eta}}^2$ como uma covariável extra e examinamos a mudança ocorrida na deviance. Como não houve uma diminuição drástica na deviance (1445,08 – 1444,78 = 0,30), evidenciamos que a função de ligação logarítmica está adequada.

Os resultados apresentados em **Tabela 3** mostram o valor da deviance do modelo em estudo e deste mesmo modelo adicionando-se $\hat{\boldsymbol{\eta}}^2$.

Tabela 3. Adequabilidade da função de ligação logarítmica para o modelo Mosquitrap_A1 - MLG

Modelo	GL	Deviance Residual	Dif. de Deviances
Y = fator(quadra) + pluv18 + tmax9 + pluv13 + tmax5 + tmin4	2195	1445,08	
Y = fator(quadra) + pluv18 + tmax9 + pluv13 + tmax5 + tmin4 + $\hat{\boldsymbol{\eta}}^2$	2194	1444,78	0,30

A contribuição de cada covariável no modelo sob pesquisa está apresentada na **Tabela 4** de análise seqüencial.

Tabela 4. ANODEV Tipo I – modelo Mosquitrap_A1 - MLG

Fonte de variação	Deviance	GL	valor *p*
Regressão	**337,98**	**104**	**0**
factor(quadra)	262,20	99	1,066e-16
pluv18	38,48	1	5,535e-10
tmax9	15,14	1	9,987e-05
pluv13	8,33	1	3,897e-03
tmax5	5,76	1	0,02
tmin4	8,07	1	4,506e-03
Resíduo	**1445,08**	**2195**	
Total	**1783,06**	**2299**	

Pela **Tabela 4**, observando o valor *p* da regressão, concluímos que há evidências, a um nível de 5% de probabilidade, que o modelo proposto está bem ajustado aos dados, havendo, ainda, a necessidade de analisar os resíduos. A **Tabela 4** mostra, ainda, que todas as variáveis independentes são significativas na presença das demais.

Para verificarmos se um termo βx no preditor linear pudesse ser melhor expresso como $\beta h(x;\lambda)$ para alguma função monótona $h(x;\lambda)$, construímos os gráficos de resíduos parciais que nos revelaram que as escalas das covariáveis são satisfatórias, uma vez que apresentaram uma tendência linear.

Em uma análise gráfica dos resíduos (r_i) *versus* valores preditos, verificamos que modelo não está adequado uma vez que os pontos não estão espalhados de forma aleatória em torno de $r_i = 0$, há uma tendência como está apresentado na **Figura 11**.

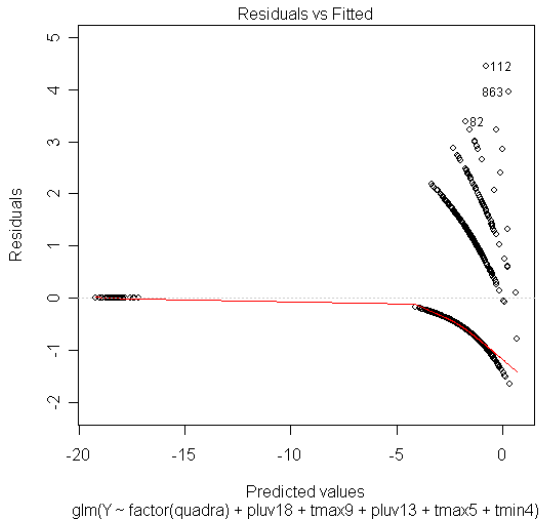


Figura 11. Gráfico de resíduos *versus* valores ajustados para o modelo Mosquitrap_A1 - MLG

Foi traçado, também, o gráfico normal de probabilidades com envelope para os componentes padronizados do desvio, conforme ilustra a **Figura 12**. Notamos que os valores extremos estão distantes de uma distribuição Poisson.

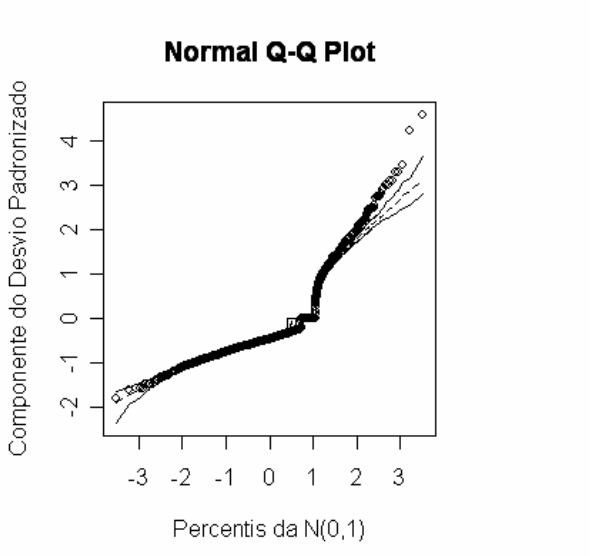


Figura 12. Gráfico normal de probabilidades com envelope simulado para o modelo Mosquitrap_A1 - MLG

Através do gráfico da distância de Cook detectamos as observações 863, 112 e 1792 como aberrantes (**Figura 13**).

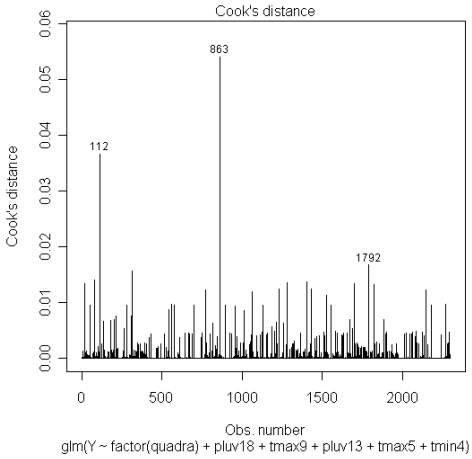


Figura 13. Gráfico da distância de Cook para o modelo Mosquitrap_A1 - MLG

Finalizando as análises de diagnósticos deste modelo, construímos os histogramas dos resíduos da deviance e de Pearson. Observamos que os gráficos não apresentaram a normalidade desejada, nos revelando, portanto, a não adequacidade do modelo ajustado.

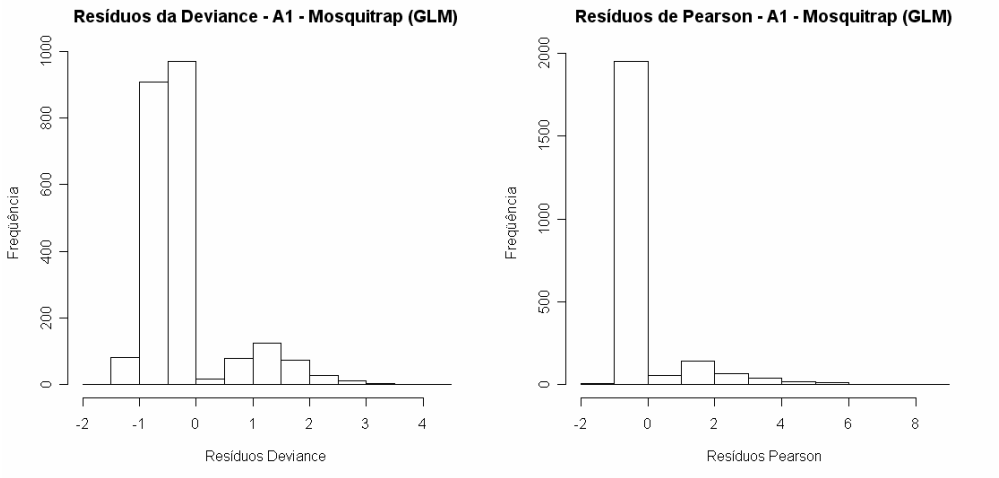


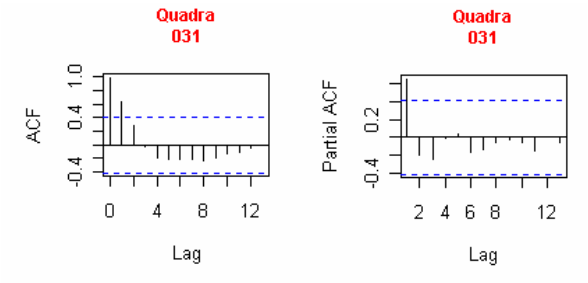
Figura 14. Histograma dos resíduos da deviance e de Pearson - modelo Mosquitrap_A1 - MLG

Como uma tentativa de encontrar um modelo que melhor se adeque aos dados, excluímos os dois maiores pontos aberrantes identificados e realizamos um novo ajuste. Este novo ajuste apresentou resultados semelhantes ao ajuste anterior com relação às análises de diagnósticos, não contribuindo, então, para que o objetivo acima citado fosse alcançado.

Realizamos, ainda, o procedimento stepwise_glm assumindo uma distribuição de Poisson e a função de ligação identidade e raiz quadrada, mas o algoritmo da função glm do R não convergiu.

4.1.2 Ajuste do modelo Mosquitrap_A1 - EEG

Constatamos, através das análises gráficas apresentadas pela **Figura 15**, que devemos assumir uma estrutura de correlação auto-regressiva de ordem 1.



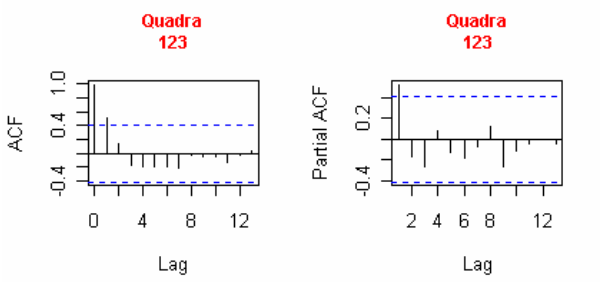


Figura 15. Correlogramas das quadras 31 e 123

O modelo obtido, as estimativas dos parâmetros e seus respectivos desvios padrão foram:

$$\hat{\eta} = \underset{(1,31)}{-0,79} + \underset{(0,002)}{0,01}\,pluv_{18}^{***} + \underset{(0,02)}{0,07}\,t\,min_{10}^{**} - \underset{(0,02)}{0,05}\,t\,max_{9}^{*} - \underset{(0,03)}{0,08}\,tmed_{2}^{**} + \underset{(0,02)}{0,06}\,t\,min_{1}^{*}$$

Prosseguindo a análise de diagnóstico deste modelo, construímos os histogramas dos resíduos ordinários e o de Pearson. Observe, através da **Figura 16**, que ambos os gráficos não apresentaram a normalidade desejada, nos indicando um ajuste inadequado.

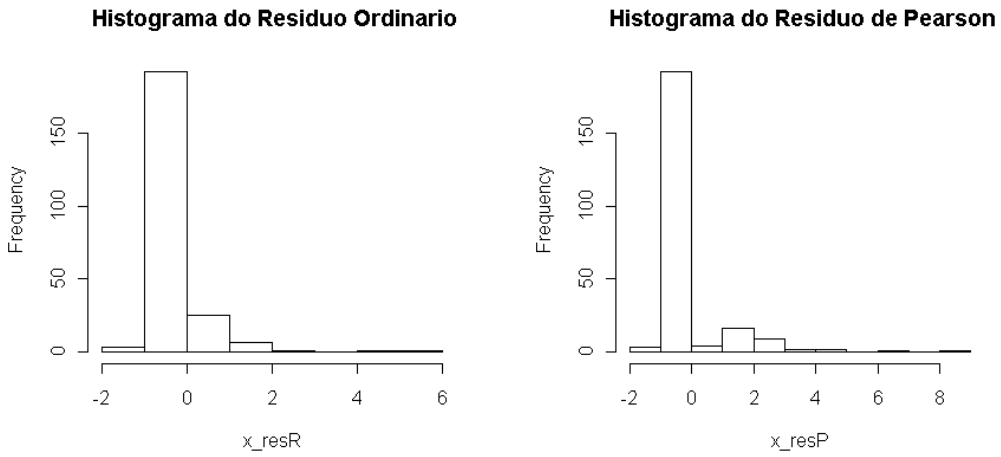


Figura 16. Histogramas dos resíduos Ordinário (esquerda) e de Pearson (direita) – modelo Mosquitrap_A1 - EEG

Os gráficos de valores preditos *versus* resíduos ordinários e de valores preditos *versus* resíduos de Pearson nos confirmaram um ajuste ruim, uma vez que os pontos não estão espalhados de forma aleatória, mas sim apresentaram uma tendência não desejada nesta análise informal mostrada na **Figura 17**.

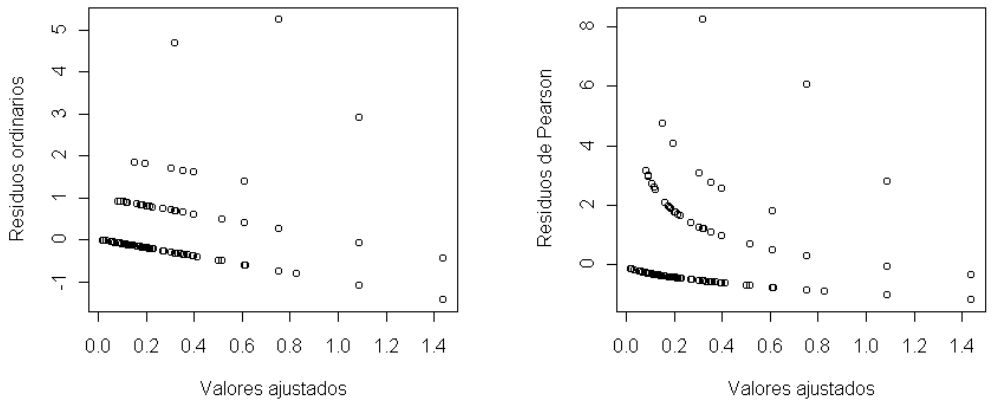


Figura 17. Gráfico de valores preditos *versus* resíduos ordinários (esquerda) e valores preditos *versus* resíduos de Pearson (direita) – modelo Mosquitrap_A1 - EEG

4.1.3 Ajuste do modelo Mosquitrap_A4 - MLG

Supondo uma distribuição de Poisson e a função de ligação logarítmica, e, ainda, considerando o domicílio como fator, o modelo obtido, as estimativas dos parâmetros e seus desvios padrão é dado por:

$$\hat{\eta} = \underset{(0,88)}{1,11} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{360\_1} + \underset{(0,453)}{0,82}D^{\bullet}_{360\_2} - \underset{(1,069)}{1,94}D^{\bullet}_{360\_3} - \underset{(1,069)}{1,94}D^{\bullet}_{360\_4} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{361\_2} - \ldots -$$

$$- \underset{(1,069)}{1,94}D^{\bullet}_{360\_4} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{400\_1} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{401\_4} - \underset{(1,069)}{1,94}D^{\bullet}_{402\_1} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{405\_1} + \ldots + \underset{(0,425)}{1,31}D^{**}_{405\_3} - \ldots -$$

$$- \underset{(1,069)}{1,94}D^{\bullet}_{406\_2} - \underset{(1,069)}{1,94}D^{\bullet}_{406\_3} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{407\_3} - \underset{(1,069)}{1,94}D^{\bullet}_{407\_4} - \underset{(1,069)}{1,94}D^{\bullet}_{447\_1} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{451\text{-}2} - \ldots -$$

$$- \underset{(1,069)}{1,94}D^{\bullet}_{452\_3} + \ldots + \underset{(0,457)}{0,76}D^{\bullet}_{458\_4} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{526\_2} - \underset{(1,069)}{1,94}D^{\bullet}_{526\_3} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{527\_4} - \ldots - \underset{(1,069)}{1,94}D^{\bullet}_{535\_1} + \ldots +$$

$$+ \underset{(0,002)}{0,01}\text{pluv}^{***}_{19} - \underset{(0,032)}{0,22}\,\text{tmin}^{***}_{15} - \underset{(0,006)}{0,04}\,\text{pluv}^{***}_{21} - \underset{(0,005)}{0,01}\,\text{pluv}^{**}_{16} + \underset{(0,006)}{0,03}\,\text{pluv}^{***}_{11} + \underset{(0,033)}{0,18}\,\text{tmin}^{***}_{9} - \underset{(0,025)}{0,08}\,\text{tmin}^{***}_{7} +$$

$$+ \underset{(0,004)}{0,01}\,\text{pluv}^{*}_{13}$$

onde nível: *** 0,001; ** 0,01; * 0,05; • 0,1

Observe que a maior parte dos domicílios significativos a um nível de 5% de probabilidade, exceto D_{360_2}, D_{405_3} e D_{458_4}, apresentaram a mesma estimativa dos parâmetros da regressão e mesmo desvio padrão.

O teste da adequabilidade da função de ligação, efetuado ao acrescentarmos $\hat{\boldsymbol{\eta}}^2$ como uma covariável extra no modelo para avaliarmos a diferença nas deviances, nos revelou

que a função de ligação logarítmica não é adequada a um nível de 5% de probabilidade, conforme ilustra a **Tabela 5**.

Tabela 5. Verificando a adequabilidade da função de ligação logarítmica para o modelo Mosquitrap_A4 - GLM

Modelo	GL	Deviance Residual	Dif. de Deviances
Y = fator(domicílio) + pluv19 + tmin15 + pluv21 + pluv16 + pluv11 + tmin9 + tmin7 + pluv13	2632	1799,93	
Y = fator(domicílio) + pluv19 + tmin15 + pluv21 + pluv16 + pluv11 + tmin9 + tmin7 + pluv13 + $\hat{\boldsymbol{\eta}}^2$	2631	1789,69	10,25

Este resultado nos levou a pesquisa de novos modelos. Assumindo uma distribuição de Poisson e função de ligação identidade e raiz quadrada tentamos ajustar novos modelos, mas o algoritmo da função glm do R não convergiu. Neste contexto, prosseguimos apresentando os resultados da análise de diagnóstico deste modelo.

A contribuição de cada covariável no modelo sob pesquisa está apresentada na **Tabela 6** de análise seqüencial.

Tabela 6. ANODEV Tipo I – modelo Mosquitrap_A4 - MLG

Fonte de variação	Deviance	GL	valor *p*
Regressão	**527,38**	**127**	**0**
factor(domicílio)	397,12	119	1,222e-31
pluv19	30,29	1	3,715e-08
tmin15	18,41	1	1,777e-05
pluv21	22,57	1	2,030e-06
pluv16	12,59	1	3,886e-04
pluv11	11,90	1	5,608e-04
tmin9	18,25	1	1,934e-05
tmin7	10,48	1	1,207e-03
pluv13	5,77	1	0,02
Resíduo	**1799,93**	**2632**	
Total	**2327,31**	**2759**	

A **Tabela 6** nos evidencia através do valor *p* da regressão, a um nível de 5% de probabilidade, que o modelo proposto está bem ajustado aos dados, havendo, ainda, a necessidade de analisar os resíduos. Observe, ainda, que todas as covariáveis são significativas na presença das demais.

Construímos, também, os gráficos dos resíduos parciais que nos revelaram que as escalas das covariáveis são satisfatórias.

Através do gráfico dos resíduos *versus* valores preditos, verificamos que modelo não está adequado pois os pontos não estão espalhados de forma aleatória em torno de $r_i = 0$, há uma tendência, como visto na **Figura 18**.

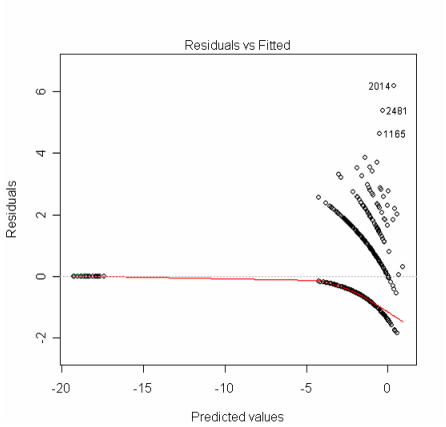


Figura 18. Gráfico dos resíduos *versus* valores preditos – modelo Mosquitrap_A4 – MLG

O gráfico normal de probabilidades com envelope simulado ilustrado pela **Figura 19** e os histogramas dos resíduos mostrados na **Figura 20** nos confirmam que o modelo não se ajustou bem aos dados.

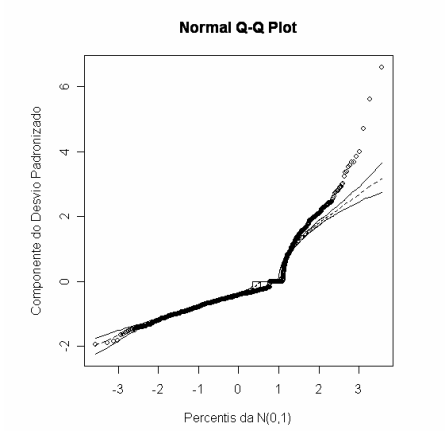


Figura 19. Gráfico normal de probabilidades – modelo Mosquitrap_A4 – MLG

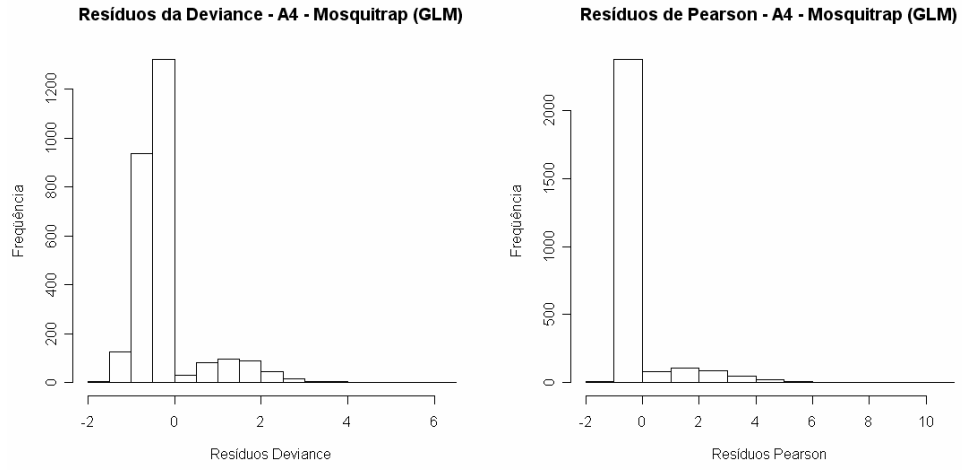


Figura 20. Histograma dos resíduos da Deviance e de Pearson – modelo Mosquitrap_A4 - MLG

O gráfico da distância de Cook (**Figura 21**) nos revelou a existência das observações 214, 2014 e 2481 como outliers.

Novamente tentamos ajustar novos modelos retirando as observações aberrantes mas, não conseguimos encontrar modelos que pudessem predizer adequadamente o valor da variável em estudo.

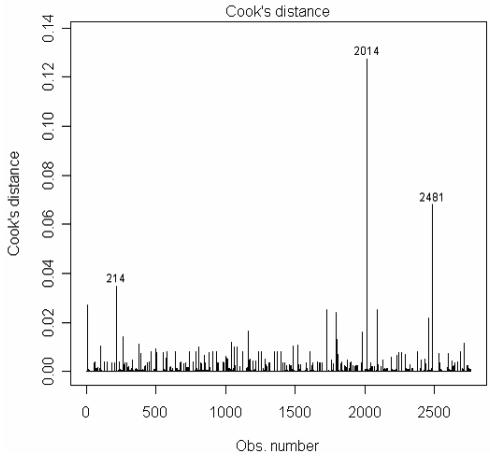


Figura 21. Gráfico da distância de Cook – modelo Mosquitrap_A4 - MLG

4.1.4 Ajuste dos modelos via abordagem de Geoestatística

Para detectar a presença da anisotropia foram construídos os gráficos do semivariograma de superfície para as 23 semanas de cada área de análise. Para a área **A1** - uma observação por quadra - estes gráficos nos revelaram que o fenômeno se espalha mais intensamente na direção Noroeste-Sudeste para todas as semanas da análise. Na área **A4** – quatro observações por quadra – a direção de anisotropia identificada para a maior parte das semanas é Norte-Sul, exceto para as semanas 6, 14, 15 e 21, onde a direção de maior continuidade espacial é concordante com a área **A1**, ou seja, ocorre mais intensamente na direção Noroeste-Sudeste.

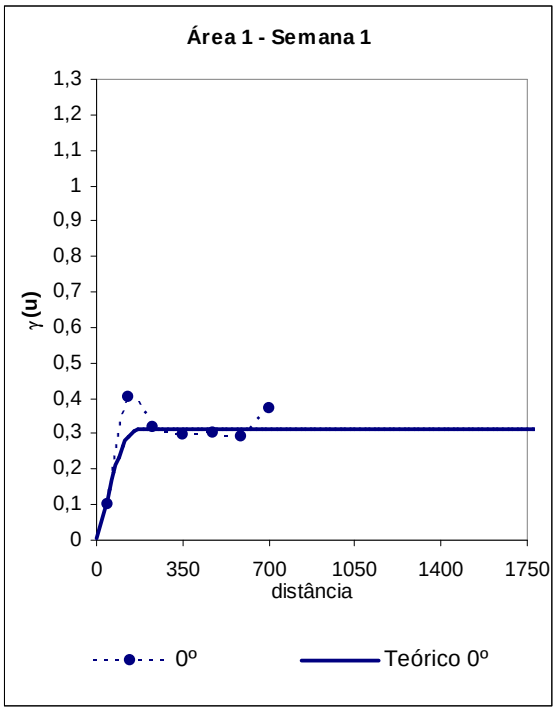
Após terem sido identificadas as direções de maior e menor continuidade espacial, prosseguimos as análises construindo os semivariogramas experimentais para cada uma destas direções. Como estamos trabalhando com amostragens irregularmente distribuídas, houve a necessidade de definir os parâmetros do *lag* e os parâmetros de direção e, desta forma, vários semivariogramas foram construídos para diferentes valores destes parâmetros. Após testarmos várias tentativas, quando conseguimos produzir um semivariograma com uma aparência desejada, escolhemos, então, os valores destes parâmetros, ilustrados em **Tabela 7**, supondo serem os mais adequados nesta primeira análise.

Tabela 7. Definição dos parâmetros do *lag* e parâmetros de direção para o cálculo do semivariograma experimental para as áreas A1 e A4

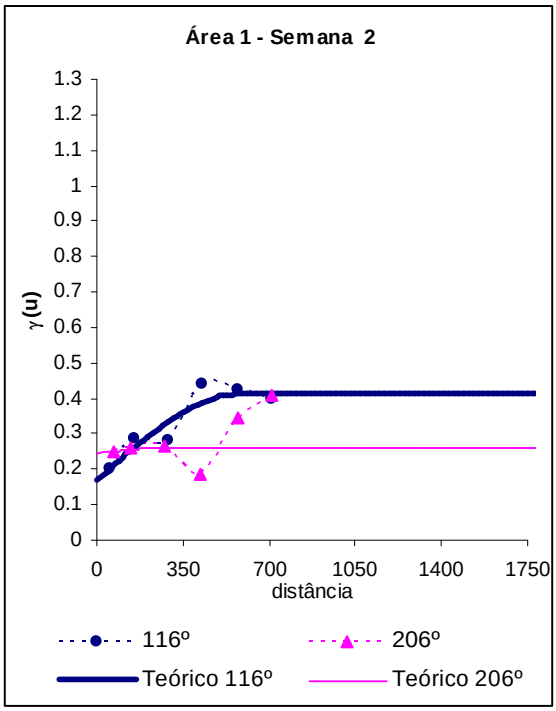
	Área A1				Área A4			
Semana	***Lag***	**Incremento**	**Intervalo de tolerância**	**Tolerância angular**	***Lag***	**Incremento**	**Intervalo de tolerância**	**Tolerância angular**
1	5	117	58,5	90	3	51	25,5	90
2	4	142	71	35	4	109	54,5	90
3	3	258	129	90	6	30	15	90
4	5	139	69,5	90	9	53	26,5	90
5	6	151	75,5	35	5	60	30	90
6	7	148	74	90	4	108	54	90
7	4	151	75,5	90	10	37	18,5	90
8	4	137	68,5	90	8	46	23	90
9	13	150	75	90	3	76	38	35
10	3	312	156	35	5	29	14,5	90
11	3	155	77,5	35	14	38	19	90
12	10	158	79	90	4	45	22,5	90
13	3	98	49	90	5	38	19	90
14	3	165	82,5	35	12	43	21,5	90
15	5	137	68,5	35	4	59	29,5	35
16	3	298	149	35	4	114	57	90
17	7	137	68,5	90	5	112	56	90
18	2	129	64,5	90	13	36	18	90
19	4	178	89	90	4	28	14	90
20	9	138	69	90	4	44	22	90
21	7	185	92,5	90	4	55	27,5	35
22	10	138	69	90	5	46	23	35
23	2	328	164	90	5	54	27	35

Nesta etapa, para a maior parte das semanas, tanto na área **A1** quanto na **A4**, não conseguimos construir um gráfico de semivariograma ideal para uma destas direções de análise e, neste contexto, assumimos um fenômeno isotrópico para estas semanas em questão. É importante destacar que na área **A4,** para as semanas que assumimos um modelo isotrópico, os gráficos do semivariograma de superfície nos indicaram uma direção de anisotropia Norte-Sul.

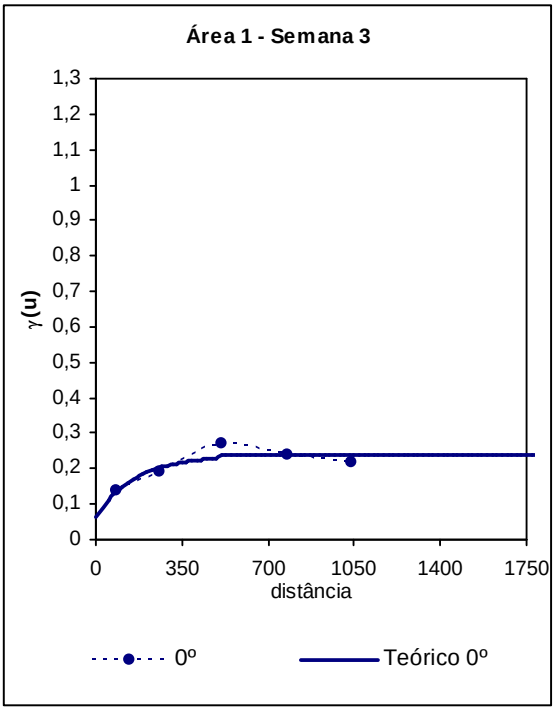
Após o ajuste de vários modelos, identificamos, tanto para a área **A1** como **A4**, a função de correlação esférica para algumas semanas e, para outras, a função de correlação exponencial como as supostas mais adequadas para expressar a estrutura de dependência espacial do fenômeno em estudo, conforme ilustra os gráficos da **Figura 22**.



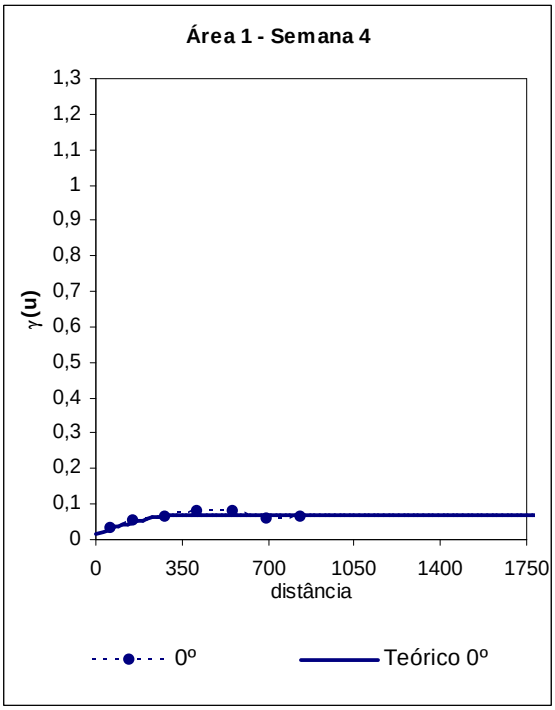
Função de correlação: esférica



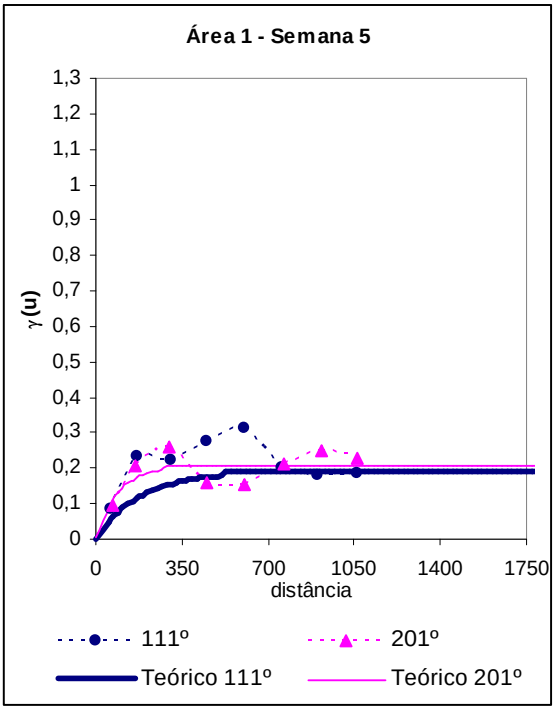
Função de correlação: esférica



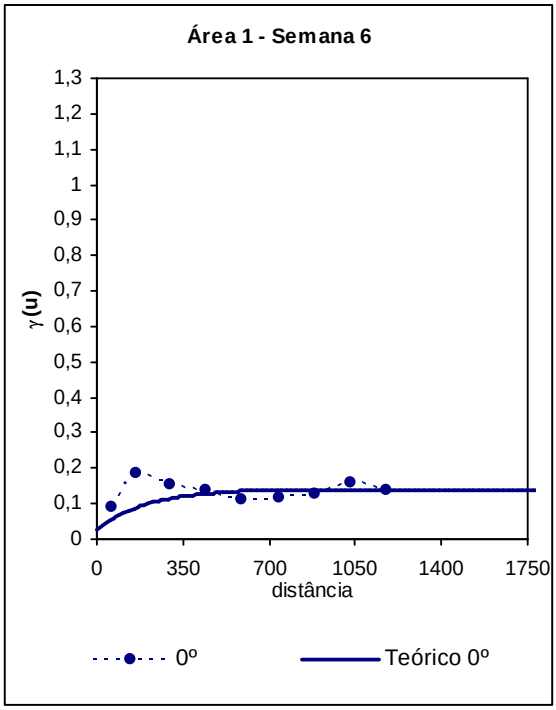
Função de correlação: exponencial



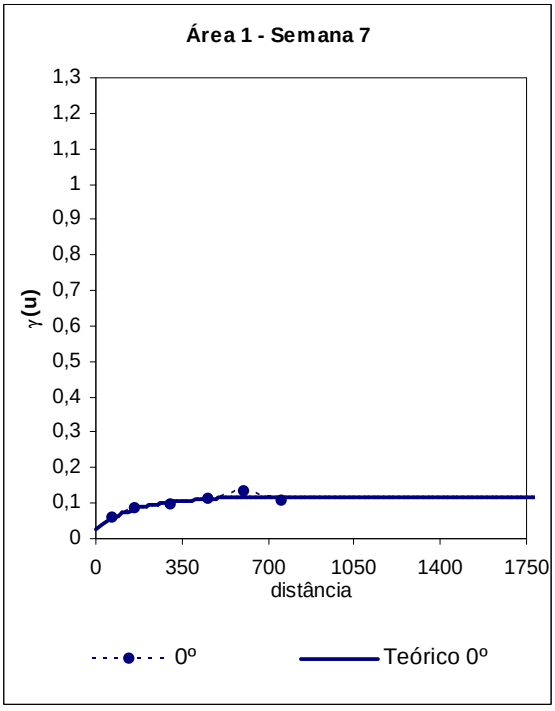
Função de correlação: esférica



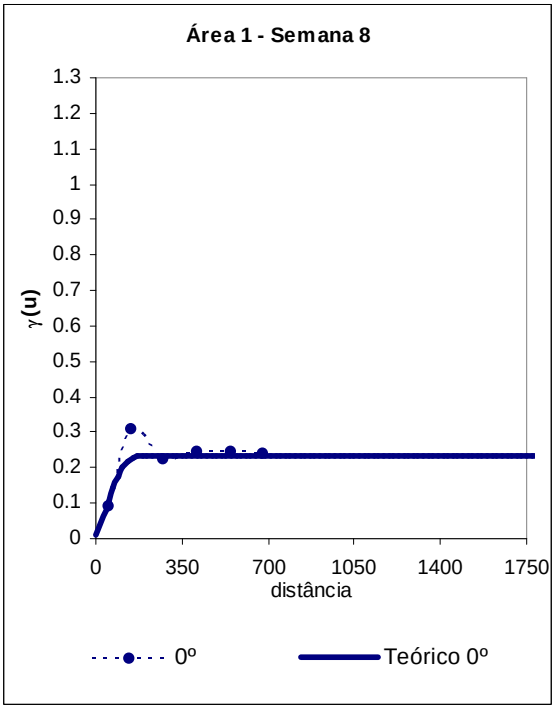
Função de correlação: exponencial



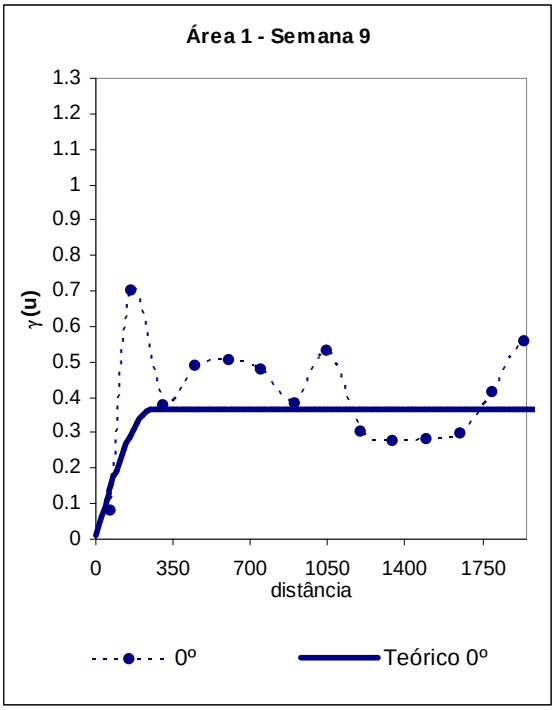
Função de correlação: exponencial



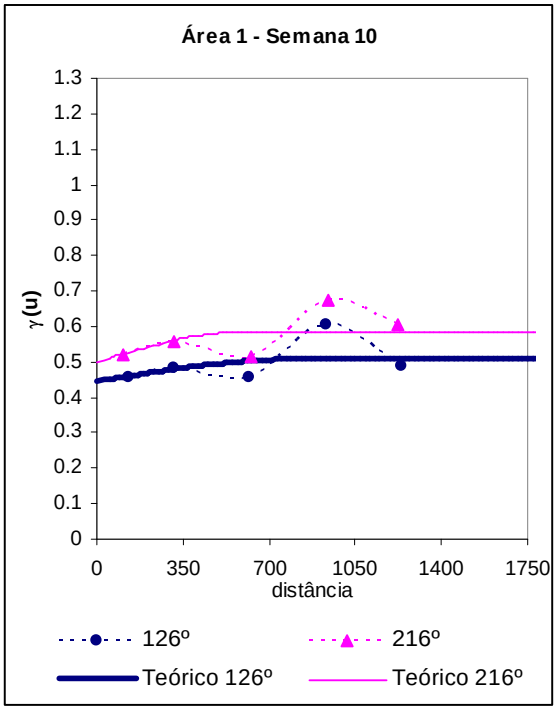
Função de correlação: exponencial



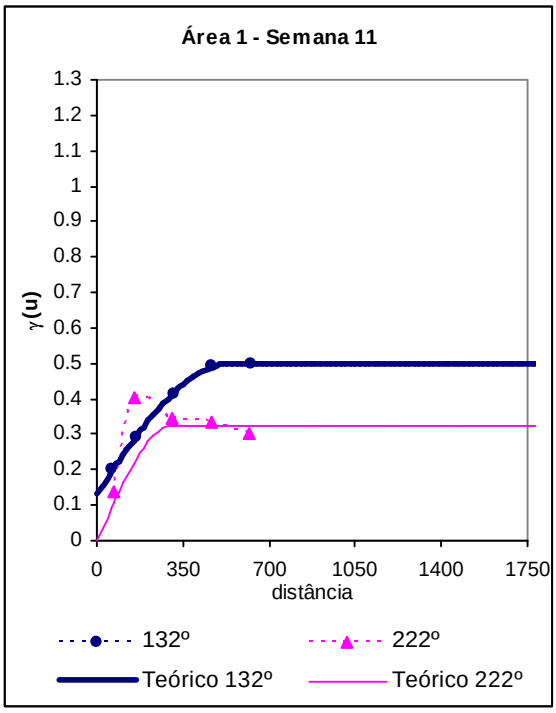
Função de correlação: esférica



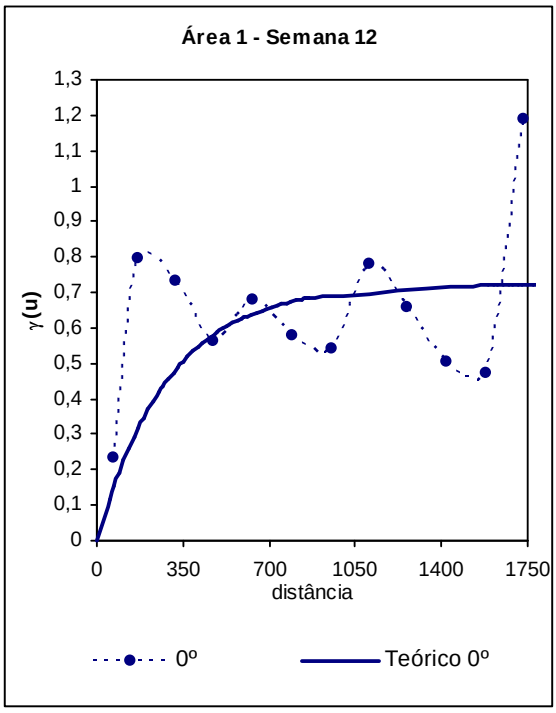
Função de correlação: esférica



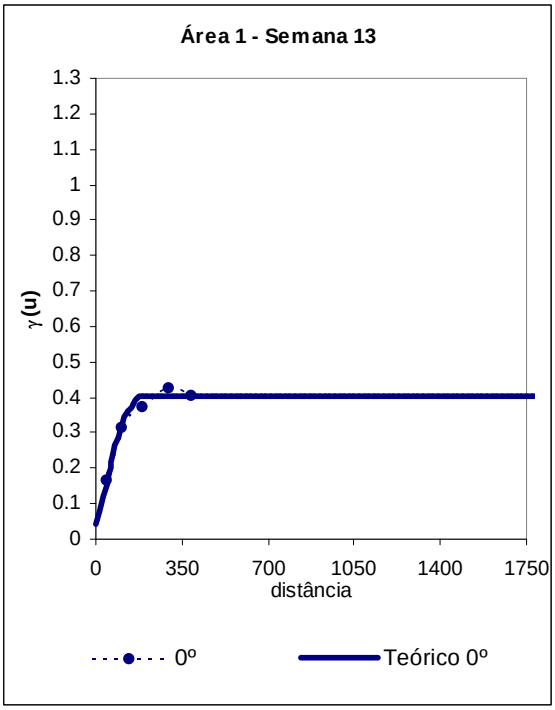
Função de correlação: esférica



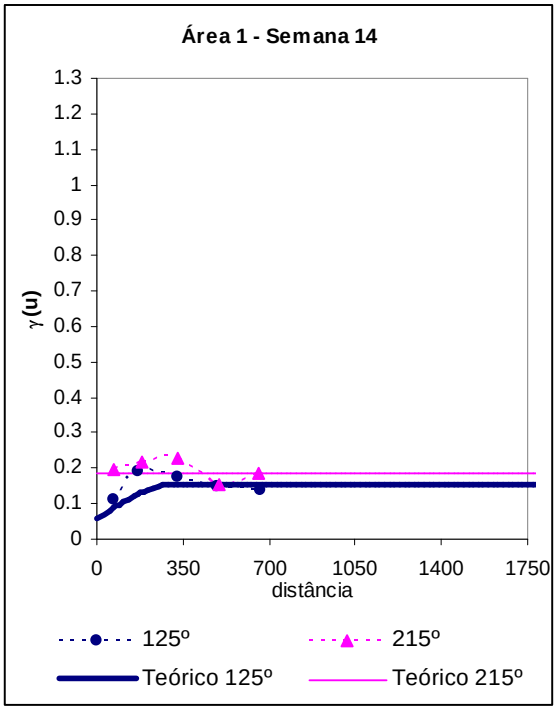
Função de correlação: esférica



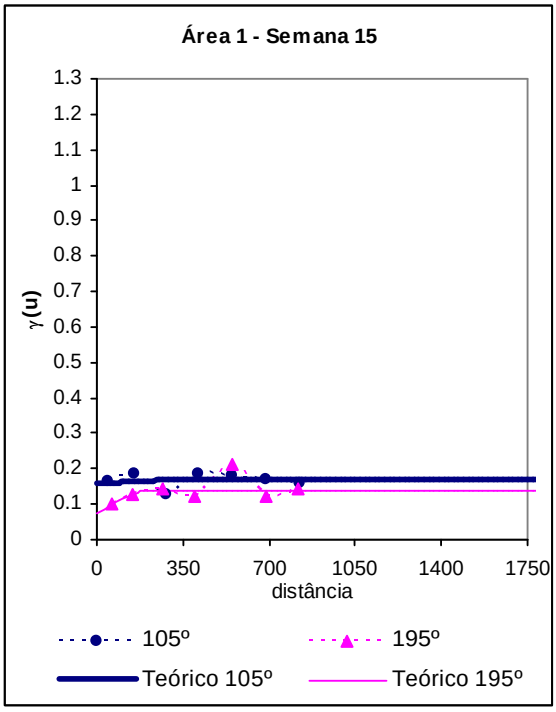
Função de correlação: exponencial



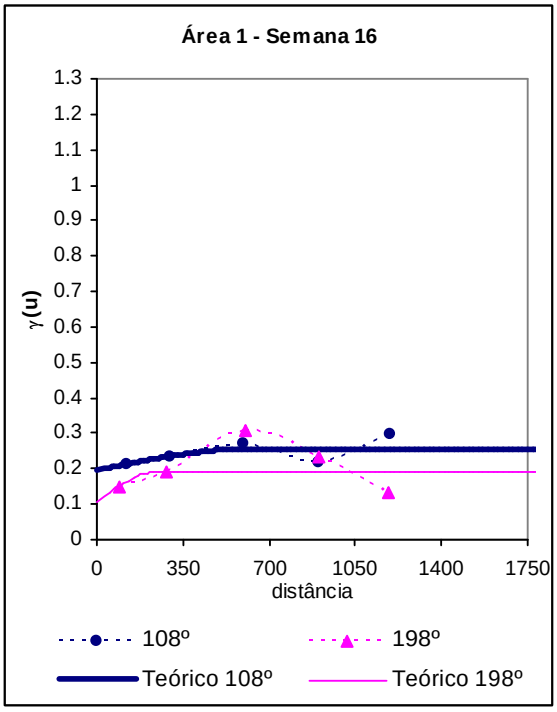
Função de correlação: esférica



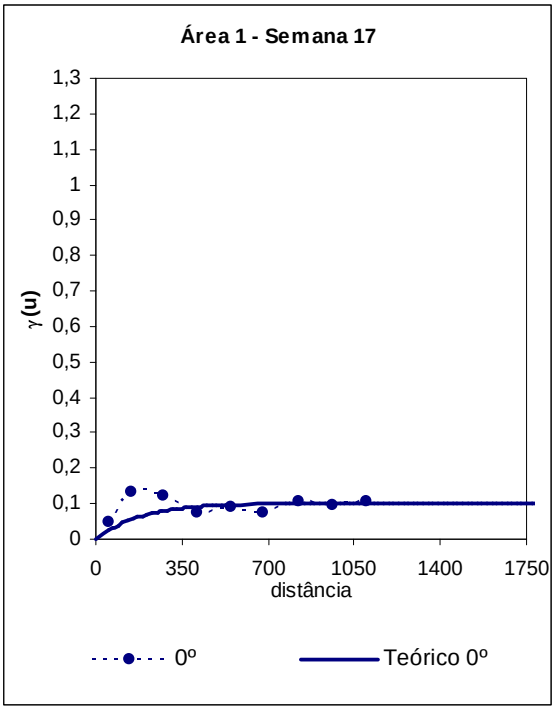
Função de correlação: esférica



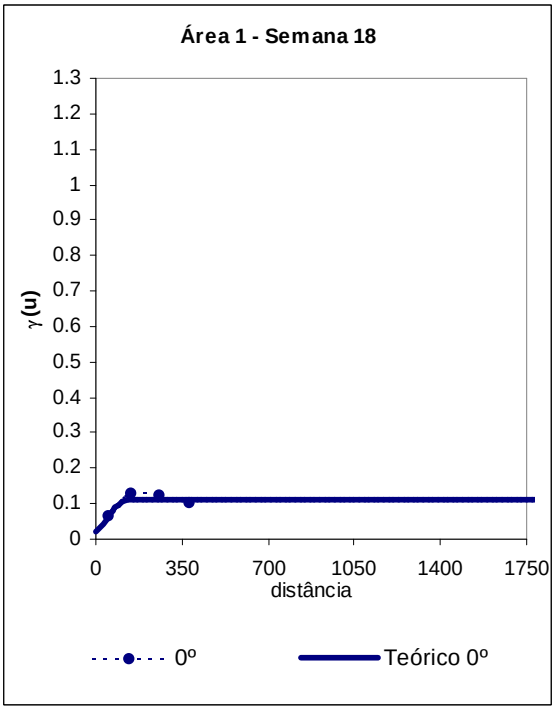
Função de correlação: esférica



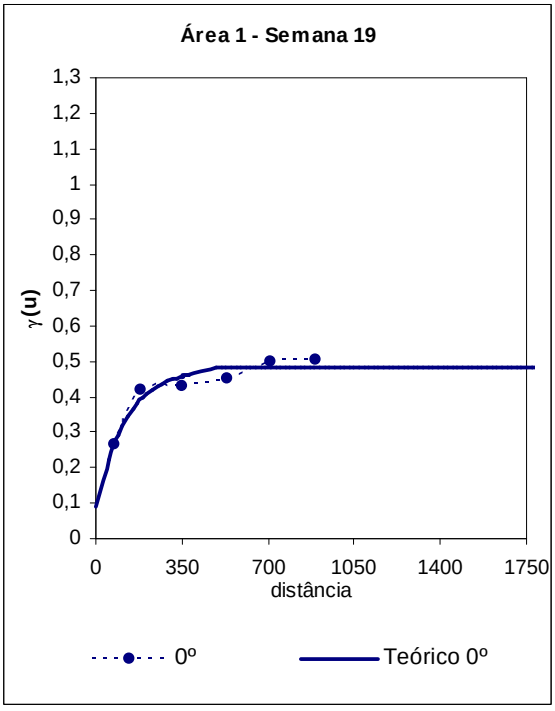
Função de correlação: esférica



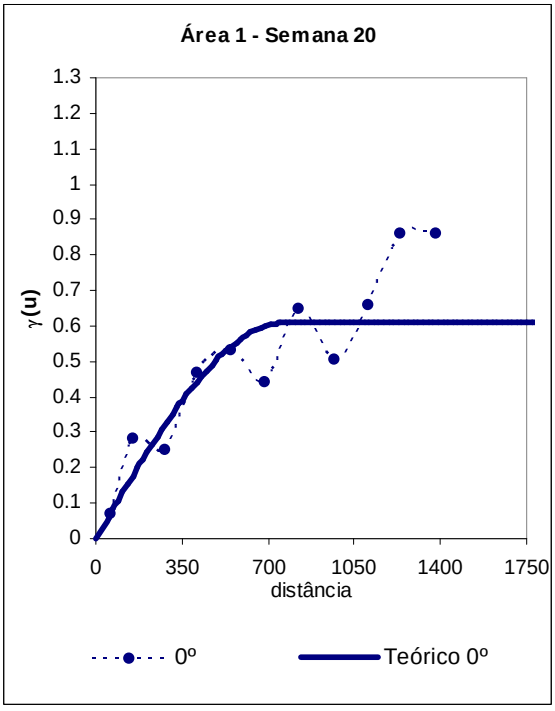
Função de correlação: exponencial



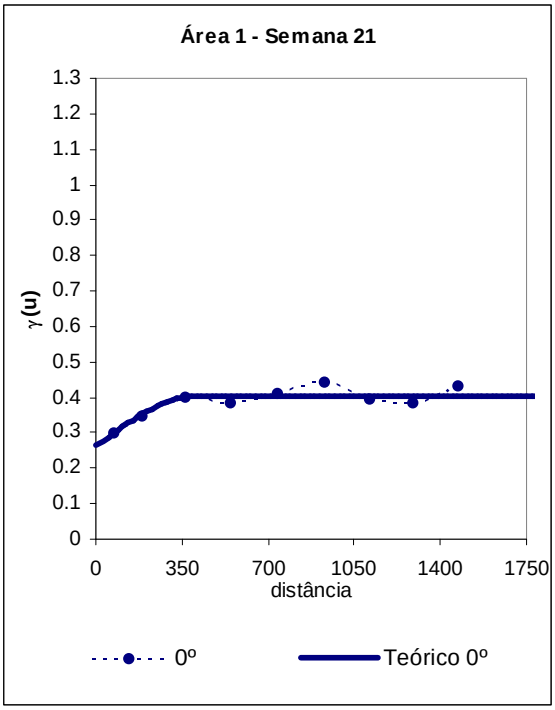
Função de correlação: esférica



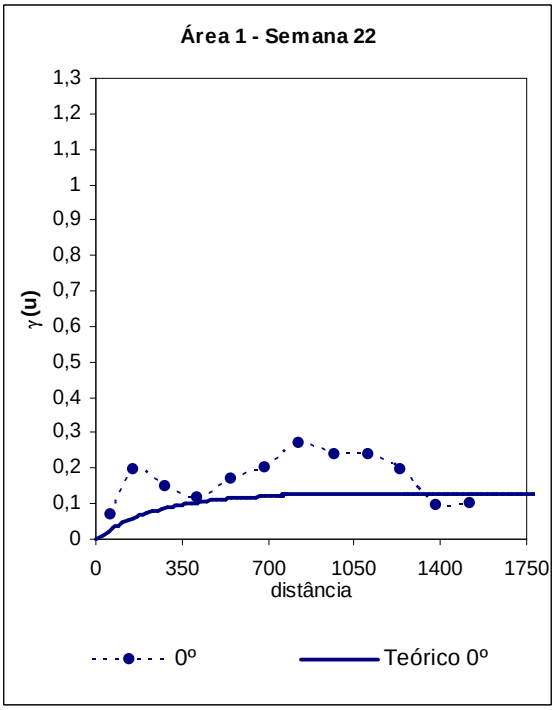
Função de correlação: exponencial



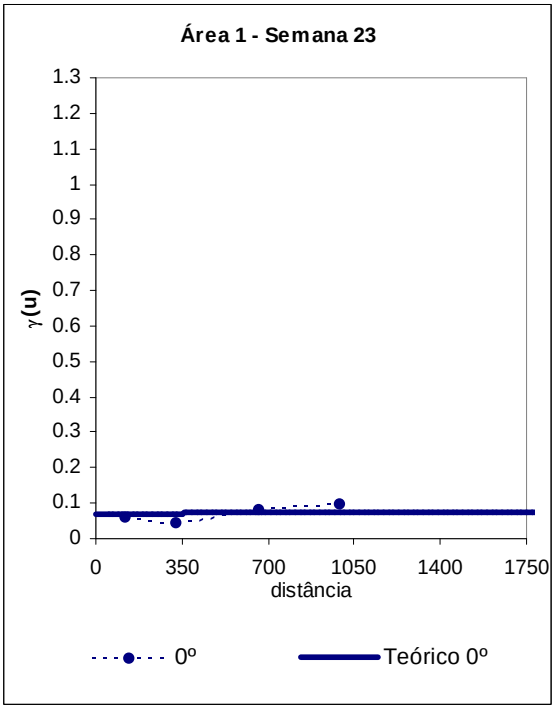
Função de correlação: esférica



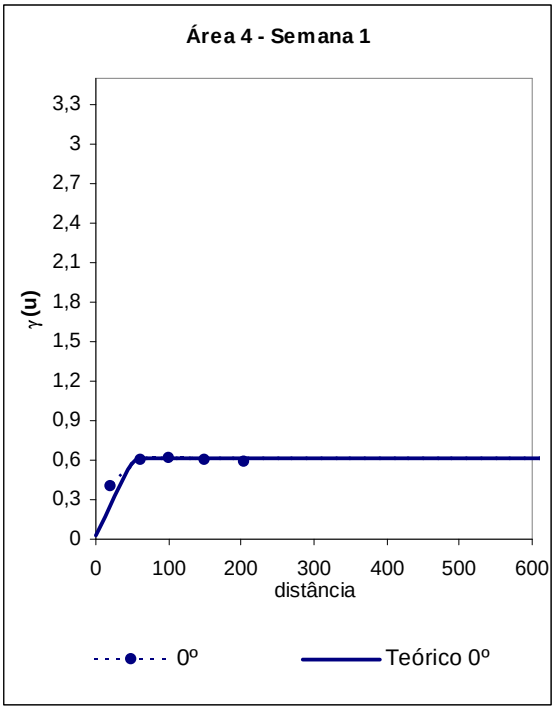
Função de correlação: esférica



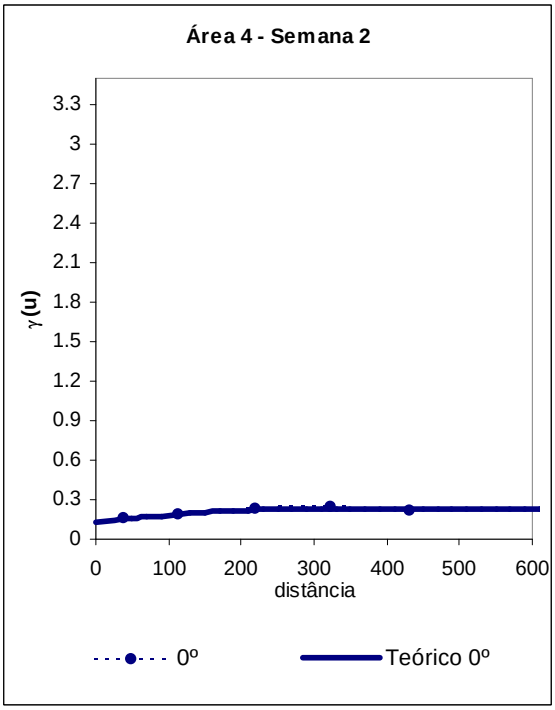
Função de correlação: exponencial



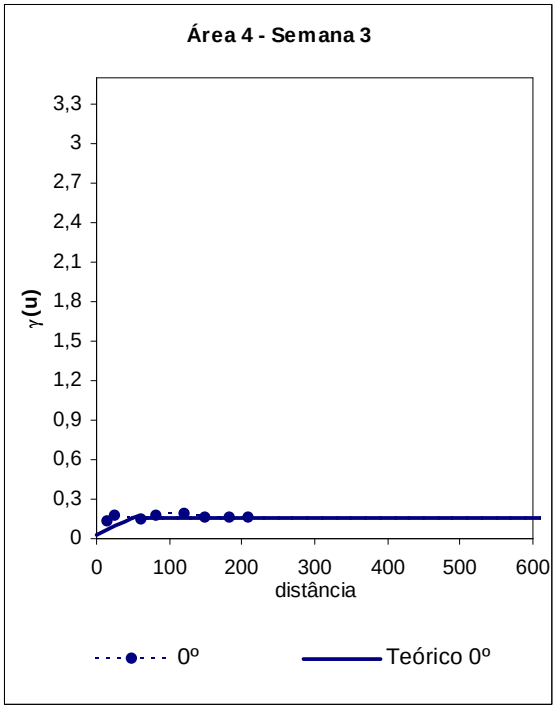
Função de correlação: esférica



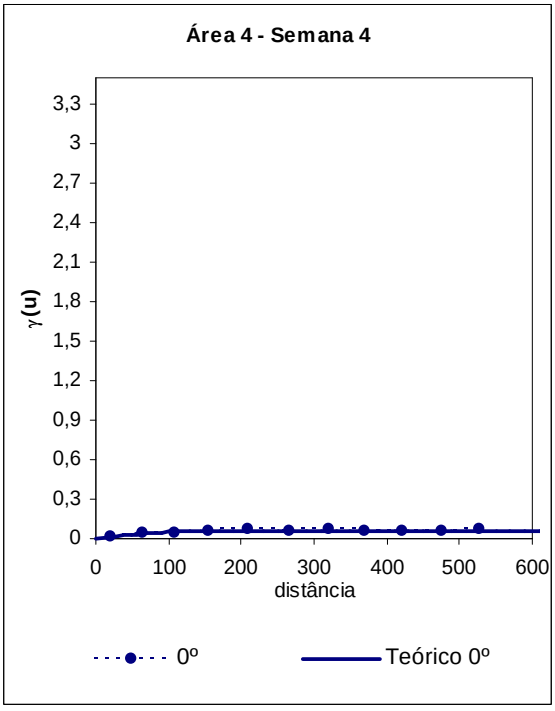
Função de correlação: exponencial



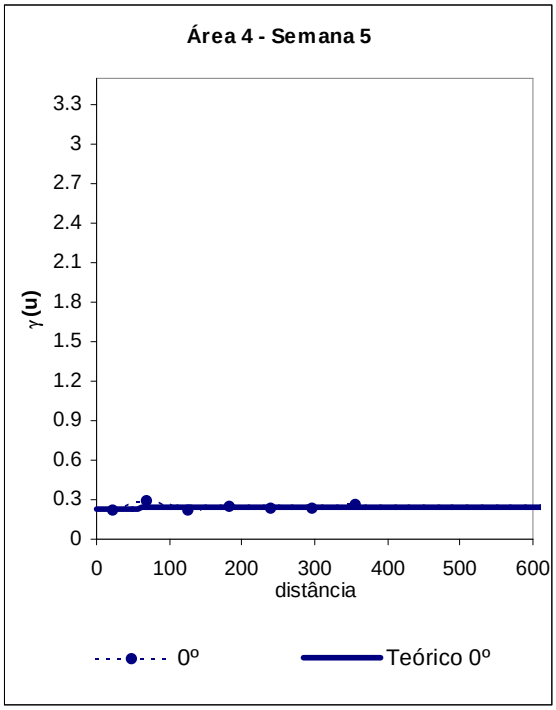
Função de correlação: esférica



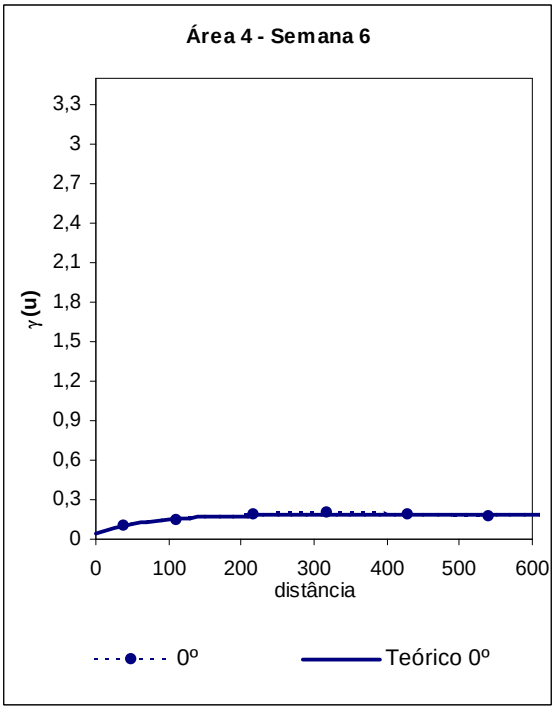
Função de correlação: exponencial



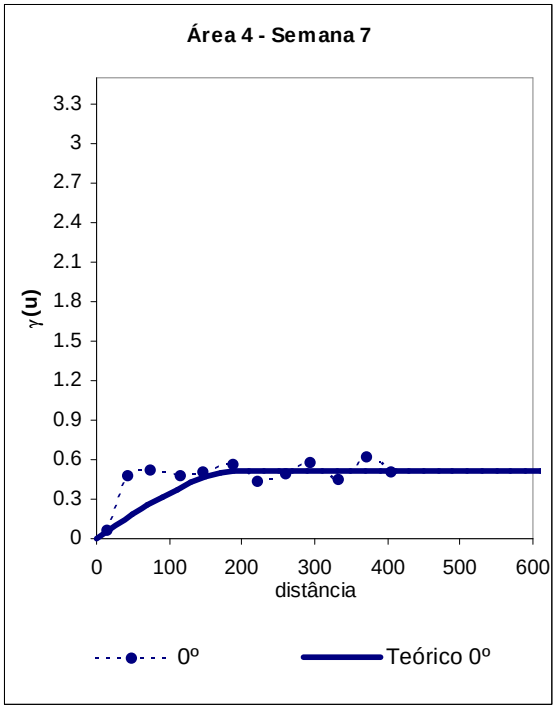
Função de correlação: exponencial



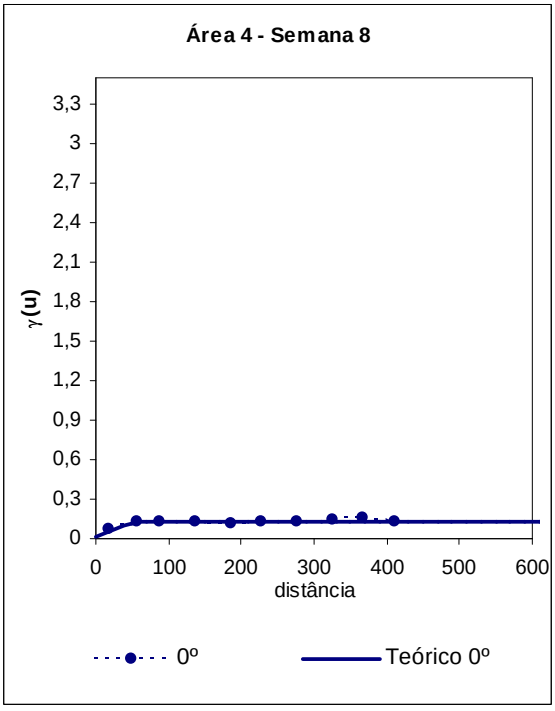
Função de correlação: esférica



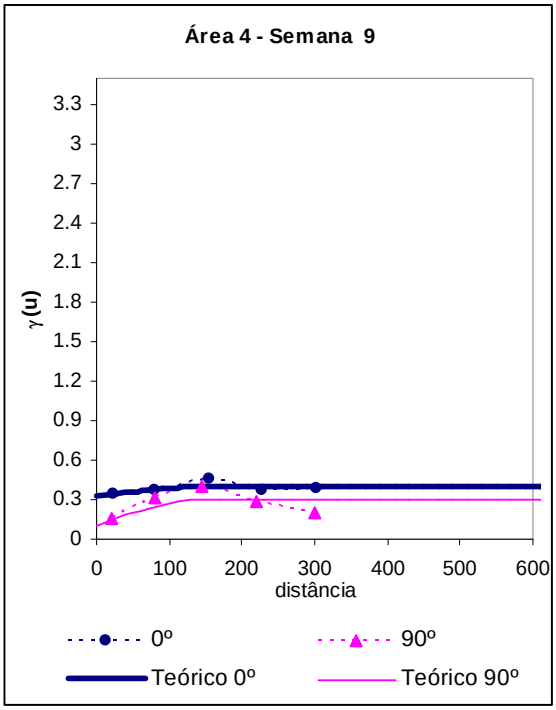
Função de correlação: exponencial



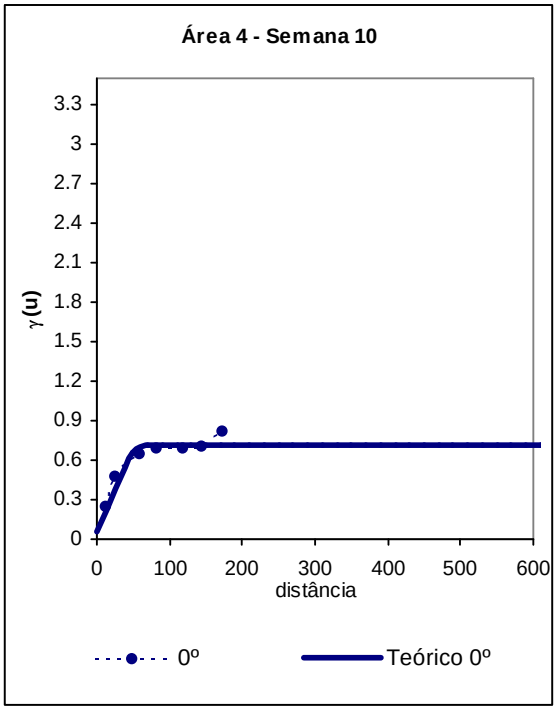
Função de correlação: esférica



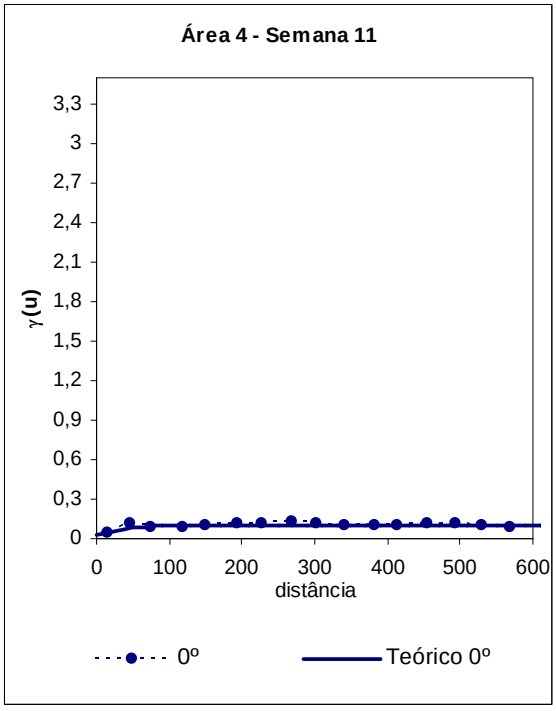
Função de correlação: exponencial



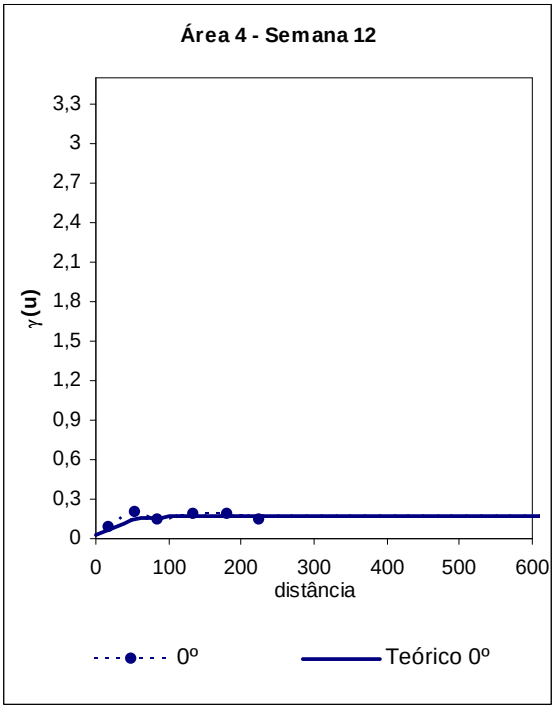
Função de correlação: esférica



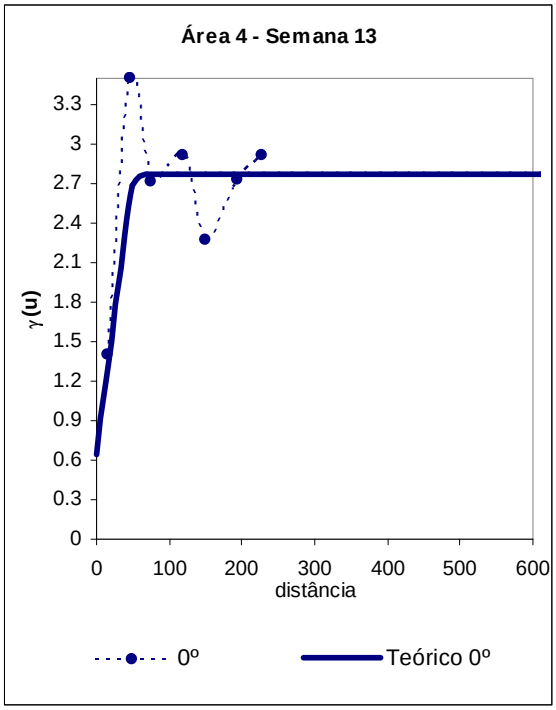
Função de correlação: esférica



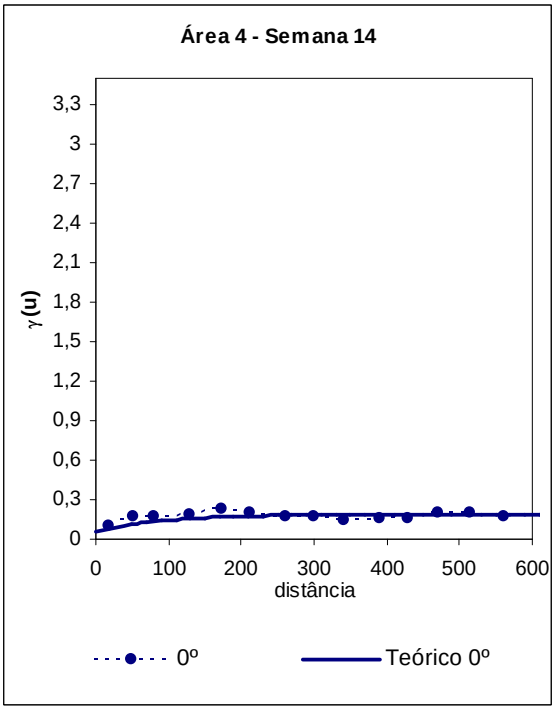
Função de correlação: exponencial



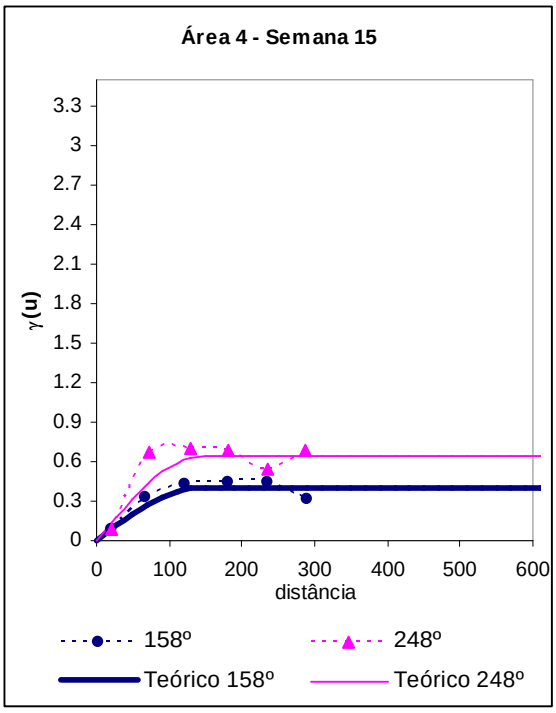
Função de correlação: exponencial



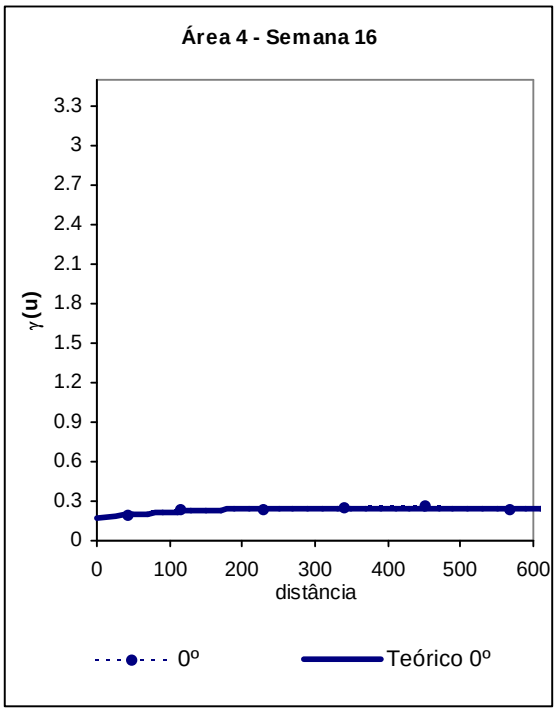
Função de correlação: esférica



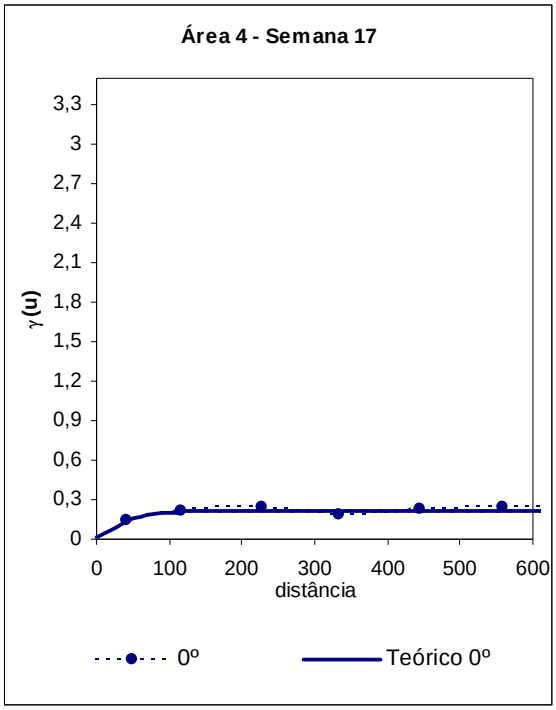
Função de correlação: exponencial



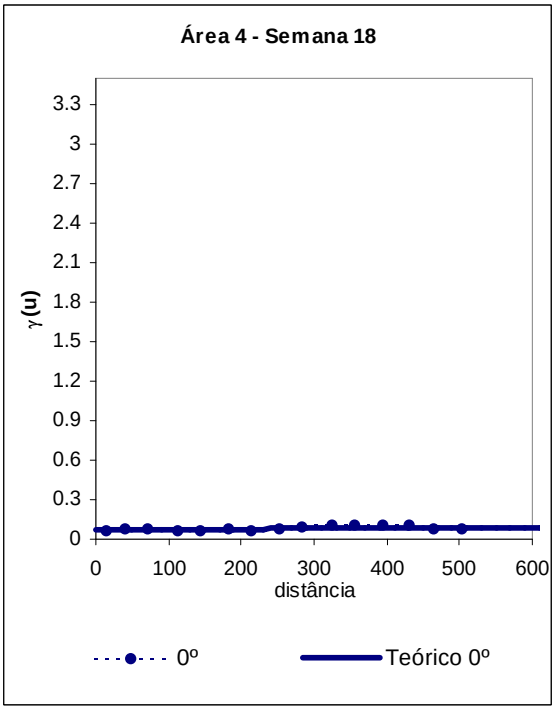
Função de correlação: esférica



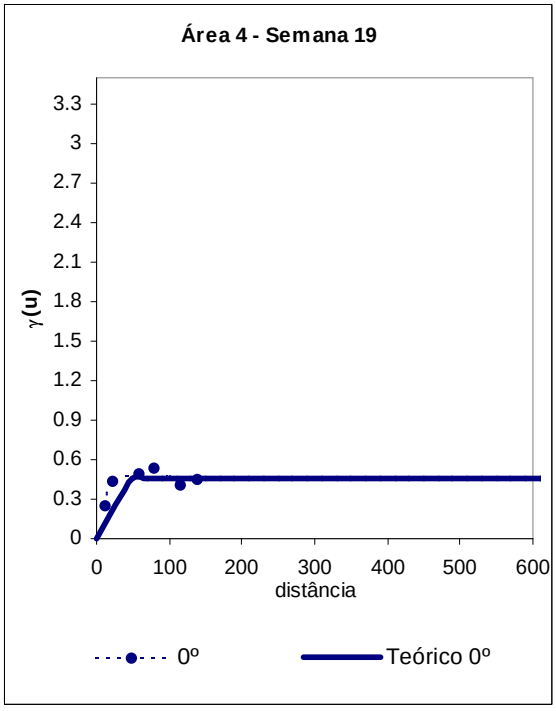
Função de correlação: esférica



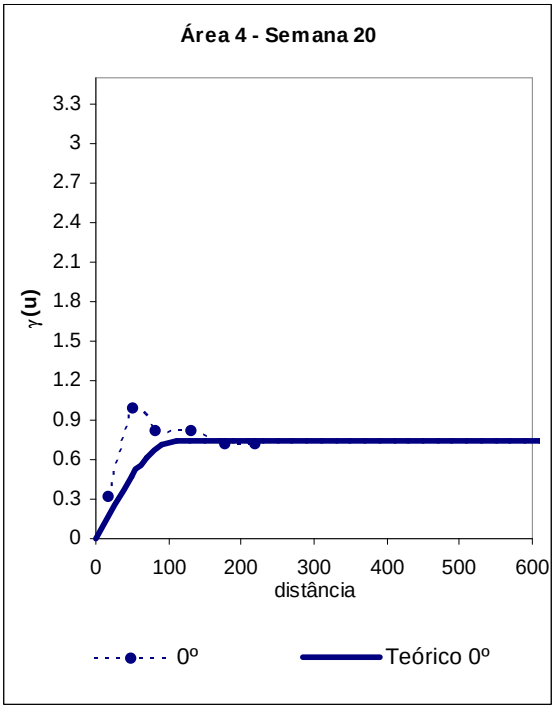
Função de correlação: exponencial



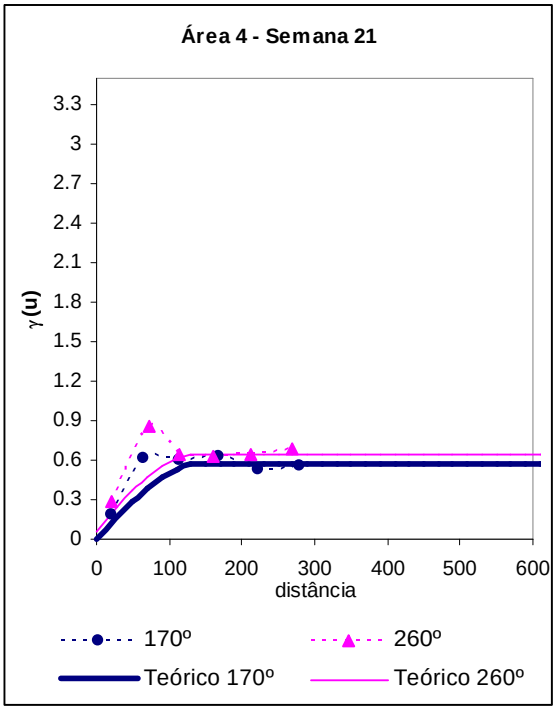
Função de correlação: esférica



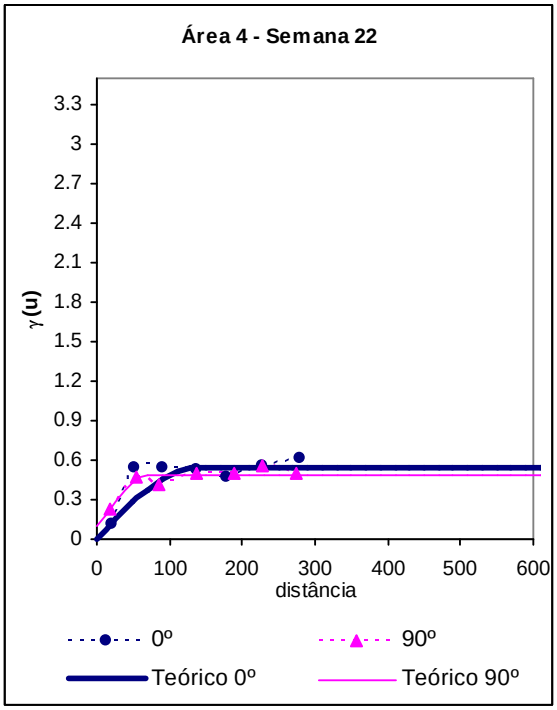
Função de correlação: esférica



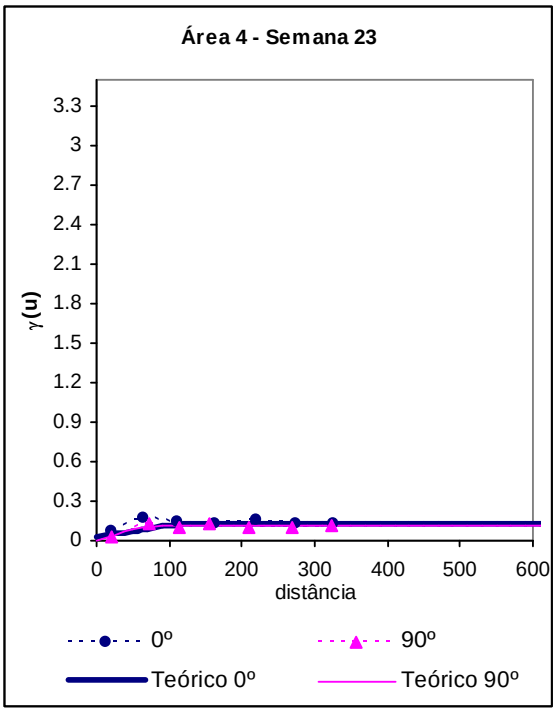
Função de correlação: esférica



Função de correlação: esférica



Função de correlação: esférica



Função de correlação: esférica

Figura 22. Semivariograma experimental e modelo ajustado para cada uma das semanas das diferentes áreas de análise, A1 e A4.

Definida a função de correlação, escolhemos os parâmetros estruturais do semivariograma teórico através do critério de seleção de Akaike. Com estes valores, calculamos

o grau de aleatoriedade presente nos dados, E = efeito pepita/patamar, e verificamos que as semanas 2, 10, 14, 15, 16, 21 e 23, área **A1**, e, semanas 2, 5, 9, 16 e 18, área **A4**, por apresentarem valores de E maiores que 0,30, apresentaram um modelo de pepita pura e, portanto, não podem ser modeladas utilizando a teoria da Geoestatística. Estes resultados estão ilustrados em **Tabela 8** e **Tabela 9**.

Tabela 8. Resultados das análises da Área A1: parâmetros estruturais do semivariograma teórico.

	Área A1 **Dados para a direção de maior e menor continuidade espacial**							
Sem.	**Ângulo**	**Akaike**	**Efeito Pepita**	**Contribuição**	**Alcance**	**Patamar**	**Função de correlação**	**E**
1	**0º**	-28,88	0,006	0,306	**165,146**	0,312	esférica	0,02
2**	**116º**	-34,855	0,168	0,245	**594,449**	0,413	esférica	**0,41**
	206º	-21,489	0,242	0,02	**297,64**	0,262		**0,92**
3	**0º**	-26,49	0,066	0,172	**496,865**	0,238	exponencial	0,28
4	**0º**	-38,73	0,016	0,052	**335,35**	0,068	esférica	0,24
5*	**111º**	-24,63	0	0,189	**529,222**	0,189	exponencial	0,00
	201º	-33,812	0,006	0,199	**274,697**	0,205		0,03
6	**0º**	-38,15	0,026	0,114	**588,942**	0,140	exponencial	0,19
7	**0º**	-34,88	0,029	0,088	**496,165**	0,117	exponencial	0,25
8	**0º**	-25,296	0,008	0,223	**161,789**	0,231	esférica	0,03
9	**0º**	-59,786	0,013	0,354	**260,841**	0,367	esférica	0,04
10**	**126º**	-21,832	0,444	0,064	**826,696**	0,508	esférica	**0,87**
	216º	-23,038	0,498	0,086	**592,915**	0,584		**0,85**
11*	**132º**	-63,414	0,135	0,363	**536,894**	0,498	esférica	0,27
	222º	-14,048	0	0,326	**310,775**	0,326		0,00
12	**0º**	-52,71	0	0,721	**868,96**	0,721	exponencial	0,00
13	**0º**	-33,159	0,042	0,359	**184,251**	0,401	esférica	0,10
14**	**125º**	-17,124	0,059	0,097	**331,009**	0,156	esférica	**0,38**
	215º	-18,39	0,184	0,004	**330,103**	0,188		**0,98**
15**	**105º**	-33,12	0,157	0,013	**411,114**	0,170	esférica	**0,92**
	195º	-32,207	0,076	0,064	**246,518**	0,14		**0,54**
16**	**108º**	-24,478	0,194	0,063	**631,954**	0,257	esférica	**0,75**
	198º	-15,567	0,108	0,083	**256,374**	**198º**		**0,57**
17	**0º**	-36,41	0	0,102	**549,676**	0,102	exponencial	0,00
18	**0º**	-21,714	0,019	0,094	**146,422**	0,113	esférica	0,17
19	**0º**	-44,28	0,092	0,389	**373,201**	0,481	exponencial	0,19
20	**0º**	-62,057	0	0,608	**765,134**	0,608	esférica	0,00
21**	**0º**	-66,213	0,263	0,141	**415,544**	0,404	esférica	**0,65**
22	**0º**	-36,19	0	0,129	**759,537**	0,129	exponencial	0,00
23**	**0º**	-11,216	0,068	0,004	**495,111**	0,072	esférica	**0,94**

* semanas onde assumimos um modelo anisotrópico;

** semanas que não podem ser modeladas via teoria da Geoestatística por apresentarem um valor para E maior que 0,30.

Tabela 9. Resultados das análises da Área A4: parâmetros estruturais do semivariograma teórico.

	Área A4 Dados para a direção de maior e menor continuidade espacial							
Sem.	**Ângulo**	**Akaike**	**Efeito Pepita**	**Contribuição**	**Alcance**	**Patamar**	**Função de correlação**	**E**
1	0º	-36,667	0,03	0,578	**51,17**	0,608	exponencial	0,04
2**	0º	-43,099	0,127	0,097	**245,73**	0,224	esférica	**0,56**
3	0º	-46,089	0,022	0,141	**27,65**	0,163	exponencial	0,13
4	0º	-66,525	0,002	0,061	**186,91**	0,063	exponencial	0,03
5**	0º	-44,113	0,231	0,009	**162,44**	0,240	esférica	**0,96**
6	0º	-45,238	0,043	0,14	**222,94**	0,183	exponencial	0,23
7	0º	-53,162	0	0,515	**203,07**	0,515	esférica	0,00
8	0º	-64,603	0,017	0,112	**61,38**	0,129	exponencial	0,13
9**	0º	-28,004	0,327	0,075	**173,24**	0,402	esférica	**0,81**
	90º	-14,852	0,102	0,198	**149,88**	0,300		**0,34**
10	0º	-42,415	0,052	0,656	**65,19**	0,708	esférica	0,07
11	0º	-95,669	0,029	0,074	**134,45**	0,103	exponencial	0,28
12	0º	-33,157	0,031	0,136	**90,90**	0,167	exponencial	0,18
13	0º	-43,424	0,649	2,123	**60,06**	2,772	esférica	0,23
14	0º	-66,837	0,054	0,135	**280,46**	0,189	exponencial	0,28
15*	158º	-39,589	0	0,4	**145,40**	0,400	esférica	0,00
	248º	-18,457	0	0,638	**143,30**	0,638		0,00
16**	0º	-44,652	0,173	0,064	**205,72**	0,237	esférica	**0,73**
17	0º	-42,789	0,012	0,208	**124,78**	0,220	exponencial	0,05
18**	0º	-73,007	0,07	0,014	**570,23**	0,084	esférica	**0,83**
19	0º	33,183	0,005	0,455	**35,73**	0,460	esférica	0,01
20	0º	-22,585	0	0,743	**109,34**	0,743	esférica	0,00
21*	170º	-20,222	0	0,57	**139,92**	0,570	esférica	0,00
	260º	-21,818	0,061	0,587	**134,92**	0,648		0,09
22*	0º	-29,904	0	0,547	**140,27**	0,547	esférica	0,00
	90º	-42,563	0,103	0,389	**76,25**	0,492		0,20
23*	0º	-26,934	0,027	0,108	**163,39**	0,135	esférica	0,20
	90º	-37,601	0,003	0,106	**81,69**	0,109		0,02

* semanas onde assumimos um modelo anisotrópico;

** semanas que não podem ser modeladas via teoria da Geoestatística por apresentarem um valor para E maior que 0,30.

Os valores obtidos para o parâmetro alcance do semivariograma teórico para a direção de maior espalhamento do fenômeno em estudo, ilustrados em **Tabela 8** e **Tabela 9**, nos revelam raios de correlação de amostras variando de 146 a 868 metros para a área **A1** e de 27 até 280 metros para a área **A4**, resultados importantes a serem considerados e analisados para que ações de controle e de prevenção mais seguras e corretas sejam efetuadas. Observe que não

estamos considerando os alcances obtidos para as semanas que não podem ser modeladas através da Geoestatística.

Após escolhermos os parâmetros estruturais do semivariograma, prosseguimos a análise definindo os modelos teóricos para as semanas que apresentaram um grau de aleatoriedade menor ou igual a 0,30, ou seja, E ≤ 0,30. Para as semanas onde conseguimos construir os semivariogramas adequados para as duas direções, apresentamos os modelos relativos às direções de máxima e mínima continuidade espacial e o modelo da anisotropia, um modelo completo, $\gamma(u)$, e consistente para qualquer distância e direção do vetor u. Quando assumimos um fenômeno isotrópico, apresentamos a expressão do único modelo ajustado. Todos estes modelos definidos são expressos a seguir:

1ª semana – A1:

$$\gamma(u) = 0{,}006 + 0{,}306\left[Sph\left(\frac{u}{165{,}15} \right) \right]$$

3ª semana – A1:

$$\gamma(u) = 0{,}066 + 0{,}172\left[Exp\left(\frac{u}{496{,}87} \right) \right]$$

4ª semana – A1:

$$\gamma(u) = 0{,}016 + 0{,}052\left[Sph\left(\frac{u}{335{,}35} \right) \right]$$

5ª semana – A1:

$$\gamma_{111^o}(u) = \lim_{a\to\infty} 0 + 0{,}189\left[Exp\left(\sqrt{\left(\frac{u_{111^o}}{529{,}22} \right)^2 + \left(\frac{u_{201^o}}{a} \right)^2} \right) \right]$$

$$\gamma_{201^o}(u) = \lim_{a\to\infty} 0{,}006 + 0{,}199\left[Exp\left(\sqrt{\left(\frac{u_{201^o}}{274{,}70} \right)^2 + \left(\frac{u_{111^o}}{a} \right)^2} \right) \right]$$

$$\begin{aligned}\gamma(u) = \lim_{\varepsilon\to 0}\lim_{a\to\infty} 0 &+ 0{,}006\left[Exp\left(\sqrt{\left(\frac{u_{111^o}}{\varepsilon} \right)^2 + \left(\frac{u_{201^o}}{274{,}70} \right)^2} \right) \right] + \\ &+ 0{,}183\left[Exp\left(\sqrt{\left(\frac{u_{111^o}}{529{,}22} \right)^2 + \left(\frac{u_{201^o}}{274{,}70} \right)^2} \right) \right] + \\ &+ 0{,}016\left[Exp\left(\sqrt{\left(\frac{u_{111^o}}{529{,}22} \right)^2 + \left(\frac{u_{201^o}}{a} \right)^2} \right) \right]\end{aligned}$$

6ª semana – A1:

$$\gamma(u) = 0{,}026 + 0{,}114\left[Exp\left(\frac{u}{588{,}94} \right) \right]$$

7ª semana – A1:

$$\gamma(u) = 0{,}029 + 0{,}088\left[Exp\left(\frac{u}{496{,}17} \right) \right]$$

8ª semana – A1:

$$\gamma(u) = 0{,}008 + 0{,}223\left[Sph\left(\frac{u}{161{,}789} \right) \right]$$

9ª semana – A1:

$$\gamma(u) = 0{,}013 + 0{,}354\left[Sph\left(\frac{u}{260{,}841} \right) \right]$$

11ª semana – A1:

$$\gamma_{132^o}(u) = \lim_{a \to \infty} 0{,}135 + 0{,}363\left[Sph\left(\sqrt{\left(\frac{u_{132^o}}{536{,}894} \right)^2 + \left(\frac{u_{222^o}}{a} \right)^2} \right) \right]$$

$$\gamma_{222^o}(u) = \lim_{a \to \infty} 0 + 0{,}362\left[Sph\left(\sqrt{\left(\frac{u_{222^o}}{310{,}775} \right)^2 + \left(\frac{u_{132^o}}{a} \right)^2} \right) \right]$$

$$\begin{aligned} \gamma(u) = \lim_{\varepsilon \to 0} \lim_{a \to \infty} 0 &+ 0{,}135\left[Sph\left(\sqrt{\left(\frac{u_{132^o}}{\varepsilon} \right)^2 + \left(\frac{u_{222^o}}{310{,}775} \right)^2} \right) \right] + \\ &+ 0{,}191\left[Sph\left(\sqrt{\left(\frac{u_{132^o}}{536{,}894} \right)^2 + \left(\frac{u_{222^o}}{310{,}775} \right)^2} \right) \right] + \\ &+ 0{,}172\left[Sph\left(\sqrt{\left(\frac{u_{132^o}}{536{,}894} \right)^2 + \left(\frac{u_{222^o}}{a} \right)^2} \right) \right] \end{aligned}$$

12ª semana – A1:

$$\gamma(u) = 0 + 0{,}721\left[Exp\left(\frac{u}{868{,}96} \right) \right]$$

13ª semana – A1:

$$\gamma(u) = 0{,}042 + 0{,}359\left[Sph\left(\frac{u}{184{,}251} \right) \right]$$

17ª semana – A1:

$$\gamma(u) = 0 + 0{,}102\left[Exp\left(\frac{u}{549{,}68} \right) \right]$$

18ª semana – A1:

$$\gamma(u) = 0{,}019 + 0{,}094\left[Sph\left(\frac{u}{146{,}422} \right) \right]$$

19ª semana – A1:

$$\gamma(u) = 0{,}092 + 0{,}389\left[Exp\left(\frac{u}{373{,}20} \right) \right]$$

20ª semana – A1:

$$\gamma(u) = 0 + 0{,}608\left[Sph\left(\frac{u}{765{,}134} \right) \right]$$

22ª semana – A1:

$$\gamma(u) = 0 + 0{,}129\left[Exp\left(\frac{u}{759{,}54} \right) \right]$$

1ª semana – A4:

$$\gamma(u) = 0{,}030 + 0{,}578\left[Exp\left(\frac{u}{51{,}17} \right) \right]$$

3ª semana – A4:

$$\gamma(u) = 0{,}022 + 0{,}141\left[Exp\left(\frac{u}{27{,}65} \right) \right]$$

4ª semana – A4:

$$\gamma(u) = 0{,}002 + 0{,}061\left[Exp\left(\frac{u}{186{,}91} \right) \right]$$

6ª semana – A4:

$$\gamma(u) = 0{,}043 + 0{,}140\left[Exp\left(\frac{u}{222{,}94} \right) \right]$$

7ª semana – A4:

$$\gamma(u) = 0 + 0{,}515\left[Sph\left(\frac{u}{203{,}070} \right) \right]$$

8ª semana – A4:

$$\gamma(u) = 0{,}017 + 0{,}112\left[Exp\left(\frac{u}{61{,}38} \right) \right]$$

10ª semana – A4:

$$\gamma(u) = 0{,}052 + 0{,}656\left[Sph\left(\frac{u}{65{,}188} \right) \right]$$

11ª semana – A4:

$$\gamma(u) = 0{,}029 + 0{,}074\left[Exp\left(\frac{u}{134{,}45} \right) \right]$$

12ª semana – A4:

$$\gamma(u) = 0{,}031 + 0{,}136\left[Exp\left(\frac{u}{90{,}90} \right) \right]$$

13ª semana – A4:

$$\gamma(u) = 0{,}649 + 2{,}123\left[Sph\left(\frac{u}{60{,}057} \right) \right]$$

14ª semana – A4:

$$\gamma(u) = 0{,}054 + 0{,}135\left[Exp\left(\frac{u}{280{,}46} \right) \right]$$

15ª semana – A4:

$$\gamma_{158^o}(u) = \lim_{a \to \infty} 0 + 0{,}400\left[Sph\left(\sqrt{\left(\frac{u_{158^o}}{145{,}403} \right)^2 + \left(\frac{u_{248^o}}{a} \right)^2} \right) \right]$$

$$\gamma_{248^o}(u) = \lim_{a \to \infty} 0 + 0{,}638\left[Sph\left(\sqrt{\left(\frac{u_{248^o}}{143{,}296} \right)^2 + \left(\frac{u_{158^o}}{a} \right)^2} \right) \right]$$

$$\begin{aligned} \gamma(u) = \lim_{\varepsilon \to 0} \lim_{a \to \infty} 0 + 0 & \left[Sph\left(\sqrt{\left(\frac{u_{158^o}}{\varepsilon} \right)^2 + \left(\frac{u_{248^o}}{143{,}296} \right)^2} \right) \right] + \\ & + 0{,}400\left[Sph\left(\sqrt{\left(\frac{u_{158^o}}{145{,}403} \right)^2 + \left(\frac{u_{248^o}}{143{,}296} \right)^2} \right) \right] + \\ & + 0{,}238\left[Sph\left(\sqrt{\left(\frac{u_{158^o}}{145{,}403} \right)^2 + \left(\frac{u_{248^o}}{a} \right)^2} \right) \right] \end{aligned}$$

17ª semana – A4:

$$\gamma(u) = 0{,}012 + 0{,}208\left[Exp\left(\frac{u}{124{,}78} \right) \right]$$

19ª semana – A4:

$$\gamma(u) = 0{,}005 + 0{,}455\left[Sph\left(\frac{u}{35{,}728} \right) \right]$$

20ª semana – A4:

$$\gamma(u) = 0 + 0{,}743\left[Sph\left(\frac{u}{109{,}344} \right) \right]$$

21ª semana – A4:

$$\gamma_{170^o}(u) = \lim_{a\to\infty} 0 + 0{,}570\left[Sph\left(\sqrt{\left(\frac{u_{170^o}}{139{,}915} \right)^2 + \left(\frac{u_{260^o}}{a} \right)^2} \right) \right]$$

$$\gamma_{260^o}(u) = \lim_{a\to\infty} 0{,}061 + 0{,}587\left[Sph\left(\sqrt{\left(\frac{u_{260^o}}{134{,}915} \right)^2 + \left(\frac{u_{170^o}}{a} \right)^2} \right) \right]$$

$$\gamma(u) = \lim_{\varepsilon\to 0}\lim_{a\to\infty} 0 + 0{,}061\left[Sph\left(\sqrt{\left(\frac{u_{170^o}}{\varepsilon} \right)^2 + \left(\frac{u_{260^o}}{134{,}915} \right)^2} \right) \right] +$$
$$+ 0{,}509\left[Sph\left(\sqrt{\left(\frac{u_{170^o}}{139{,}915} \right)^2 + \left(\frac{u_{260^o}}{134{,}915} \right)^2} \right) \right] +$$
$$+ 0{,}078\left[Sph\left(\sqrt{\left(\frac{u_{170^o}}{139{,}915} \right)^2 + \left(\frac{u_{260^o}}{a} \right)^2} \right) \right]$$

22ª semana – A4:

$$\gamma_{0^o}(u) = \lim_{a\to\infty} 0 + 0{,}547\left[Sph\left(\sqrt{\left(\frac{u_{0^o}}{140{,}265} \right)^2 + \left(\frac{u_{90^o}}{a} \right)^2} \right) \right]$$

$$\gamma_{90^o}(u) = \lim_{a\to\infty} 0{,}103 + 0{,}389\left[Sph\left(\sqrt{\left(\frac{u_{90^o}}{76{,}251} \right)^2 + \left(\frac{u_{0^o}}{a} \right)^2} \right) \right]$$

$$\gamma(u) = \lim_{\varepsilon \to 0} \lim_{a \to \infty} 0 + 0{,}103 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{\varepsilon} \right)^2 + \left(\frac{u_{90^\circ}}{76{,}251} \right)^2} \right) \right] +$$
$$+ 0{,}389 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{140{,}265} \right)^2 + \left(\frac{u_{90^\circ}}{76{,}251} \right)^2} \right) \right] +$$
$$+ 0{,}055 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{140{,}265} \right)^2 + \left(\frac{u_{90^\circ}}{a} \right)^2} \right) \right]$$

23ª semana – A4:

$$\gamma_{0^\circ}(u) = \lim_{a \to \infty} 0{,}027 + 0{,}108 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{163{,}386} \right)^2 + \left(\frac{u_{90^\circ}}{a} \right)^2} \right) \right]$$

$$\gamma_{90^\circ}(u) = \lim_{a \to \infty} 0{,}003 + 0{,}106 \left[Sph\left(\sqrt{\left(\frac{u_{90^\circ}}{81{,}686} \right)^2 + \left(\frac{u_{0^\circ}}{a} \right)^2} \right) \right]$$

$$\gamma(u) = \lim_{\varepsilon \to 0} \lim_{a \to \infty} 0{,}003 + 0{,}024 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{\varepsilon} \right)^2 + \left(\frac{u_{90^\circ}}{81{,}686} \right)^2} \right) \right] +$$
$$+ 0{,}082 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{163{,}386} \right)^2 + \left(\frac{u_{90^\circ}}{81{,}686} \right)^2} \right) \right] +$$
$$+ 0{,}026 \left[Sph\left(\sqrt{\left(\frac{u_{0^\circ}}{163{,}386} \right)^2 + \left(\frac{u_{90^\circ}}{a} \right)^2} \right) \right]$$

Como última etapa, executamos o processo de validação destes modelos para avaliar o grau de incerteza ao utilizarmos os parâmetros escolhidos. Obtivemos, então, as informações ilustradas nas tabelas **Tabela 10** e **Tabela 11** referentes ao erro destas estimativas.

Tabela 10. Informações referentes ao erro das estimativas – área A1

Área A1								
Sem.	**Média**	**Variância**	**Desvio Padrão**	**Coeficiente de Variação**	**Coeficiente de Assimetria**	**Coeficiente de Curtose**	**Valor Mínimo**	**Valor Máximo**
1	-0,0070	0,1370	0,3700	-52,6540	-3,3460	24,8210	-2,5480	1,0280
3	0,0110	0,0640	0,2530	23,6180	-1,3470	6,8320	-1,0300	0,7240
4	-0,0050	0,0250	0,1570	-32,2210	-2,9800	19,4710	-0,8840	0,5420
5	-0,0160	0,1190	0,3450	-21,2450	3,0410	17,3330	-1,8640	0,6900
6	-0,0230	0,0900	0,3000	-13,2220	-3,6520	22,0770	-1,9760	0,8000
7	0,0070	0,0480	0,2190	29,9370	-2,0020	11,1670	-0,9590	0,6260
8	0,0020	0,0880	0,2960	170,0820	-1,0280	12,0990	-1,1420	1,2850
9	-0,0400	0,3470	0,5890	-14,6270	-5,7730	48,4250	-4,8860	1,7620
11	0,0050	0,2040	0,4520	86,5560	-1,8010	8,2490	-2,0360	1,1660
12	0,0580	0,2240	0,4740	8,0980	0,8900	11,3260	-2,0060	2,0200
13	0,0130	0,1640	0,4050	31,5630	-4,2010	32,1090	-3,0000	0,9220
17	-0,0020	0,0490	0,2210	-110,2920	-6,9470	68,6670	-2,0070	0,7300
18	0,0010	0,0640	0,2540	208,2030	-4,9610	41,7030	-2,0000	0,7190
19	0,0060	0,1670	0,4080	68,9950	-0,6770	11,2670	-1,9510	1,7760
20	0,0060	0,0330	0,1830	31,5640	-0,0520	21,1180	-0,9190	1,0420
22	0,0070	0,0470	0,2160	30,5730	-1,6980	15,6890	-1,0000	0,7670

Tabela 11. Informações referentes ao erro das estimativas – área A4

Área A4								
Sem.	**Média**	**Variância**	**Desvio Padrão**	**Coeficiente de Variação**	**Coeficiente de Assimetria**	**Coeficiente de Curtose**	**Valor Mínimo**	**Valor Máximo**
1	-0,0010	0,0890	0,2980	-366,3400	-4,0260	31,2350	-2,2730	0,9310
3	0,0150	0,0240	0,1560	10,1830	-2,4190	16,0680	-0,8740	0,4860
4	0,0020	0,0100	0,1020	41,0570	-6,9940	72,9640	-0,9830	0,3030
6	0,0220	0,0670	0,2580	11,9260	0,2900	16,1090	-1,2120	1,4420
7	0,0340	0,1530	0,3920	11,6860	-0,2620	12,7780	-1,9780	1,6750
8	0,0010	0,0450	0,2130	216,6420	-5,8870	50,8620	-1,8410	0,5220
10	0,0570	0,1190	0,3440	6,0340	1,3680	11,7690	-0,9150	1,9130
11	-0,0020	0,0320	0,1800	-111,5240	-3,4340	18,9910	-1,0000	0,4700
12	-0,0020	0,0370	0,1930	-97,4080	-3,3740	18,2570	-1,0000	0,4100
13	0,0280	0,5400	0,7350	25,8030	-1,7640	14,9140	-3,6340	2,9820
14	0,0160	0,0660	0,2580	16,1470	-1,4890	8,2480	-0,9980	0,8140
15	0,0120	0,0410	0,2020	17,5260	4,4820	51,7160	-1,0220	1,7440
17	-0,0300	0,0750	0,2740	-9,0710	-4,7520	29,0950	-1,7860	0,6240
19	-0,0040	0,0850	0,2920	-69,3830	-4,3030	27,3730	-1,9260	0,5850
20	-0,0110	0,1590	0,3980	-34,9860	-2,5070	20,5090	-2,0740	1,5520
21	-0,0190	0,2660	0,5160	-26,5000	-3,7640	31,5830	-4,0000	1,4320
22	0,0170	0,0930	0,3060	17,6120	-0,1970	8,2500	-0,9940	1,2440
23	-0,0220	0,0790	0,2810	-12,5500	-4,2350	25,3110	-1,9940	0,4120

Analisando os valores das **Tabela 10** e **Tabela 11**, verificamos que as estatísticas apresentam valores aceitáveis dentro das suposições impostas ao erro das estimativas. Vale destacar que os modelos com maiores desvios padrão são os das semanas 9, 11, 12, 13 e 19 para a área **A1** e, semanas 7, 13, 20 e 21, para a área **A4**, gerando, como conseqüência, uma superestimação ou subestimação dos valores amostrais. Provavelmente isto ocorre devido ao excesso de zeros existentes nas duas áreas de amostragem, ou, de amostragens inadequadas ou com poucas observações, ou, ainda, do modelo adotado que pode não estar representando de forma adequada a variabilidade espacial do número de fêmeas *Aedes* capturadas por armadilhas adulticidas.

Foram também construídos os histogramas dos erros, o diagrama dos valores observados *versus* estimados e a distribuição espacial do erro. Todas estas representações gráficas nos confirmaram os resultados acima relatados.

4.2 Resultados dos modelos elaborados para estudar as variáveis respostas X e Z, os alcances de maior e menor espalhamento do número fêmeas *Aedes* capturadas em armadilhas adulticidas – área A1

Os histogramas construídos para as variáveis aleatórias **X** e **Z** nos revelaram que podemos supor, inicialmente, que ambas possuem uma distribuição Gama, conforme ilustra a **Figura 23**.

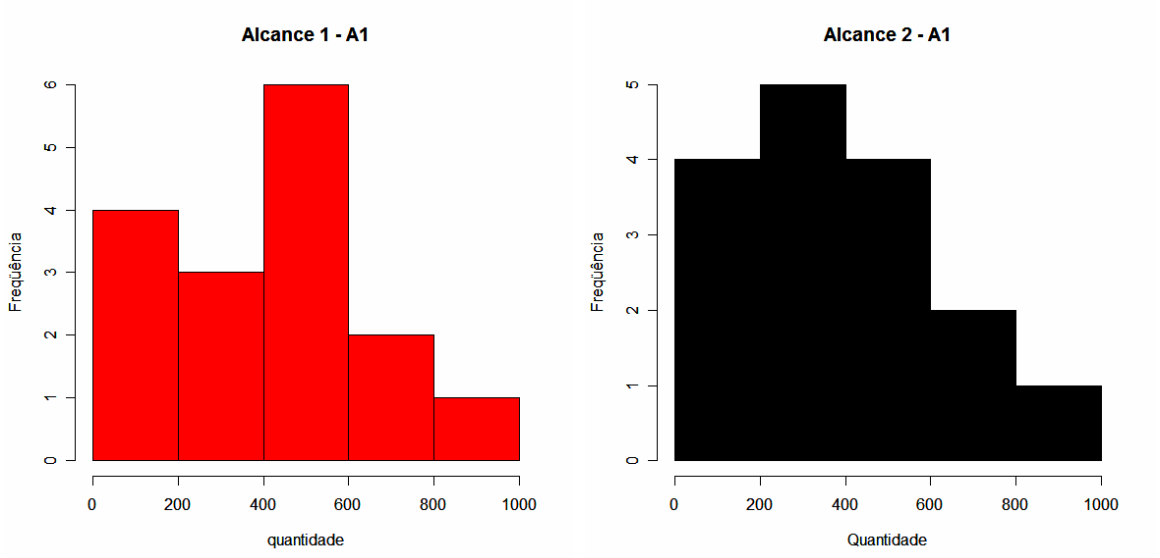


Figura 23. Histogramas das amostras X, à esquerda, e Z, à direita.

As análises via correlogramas não nos revelaram a existência de uma correlação temporal para as variáveis aleatórias em questão, por isso efetuamos modelagens via teoria de MLG (**Figura 24**).

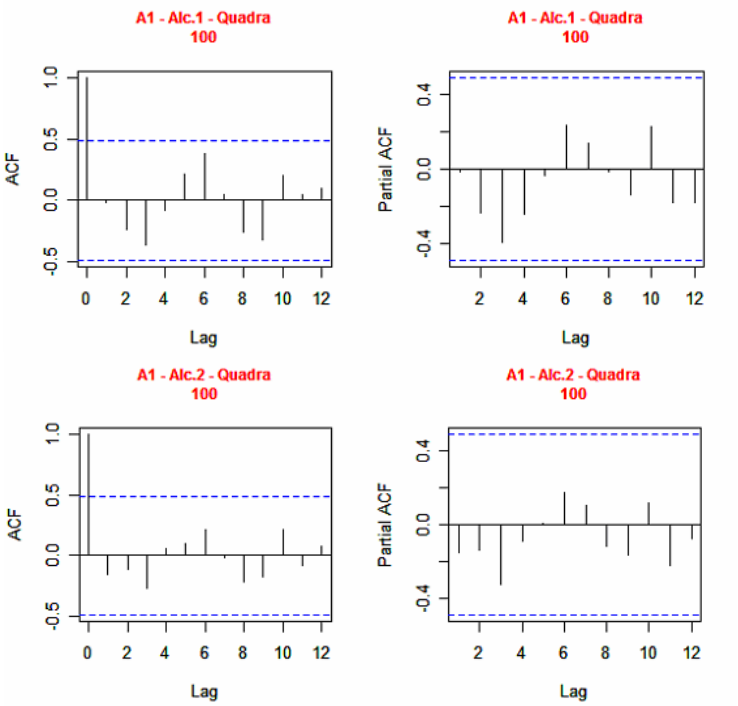


Figura 24. Correlogramas para X, alcance1 e Z, alcance2 – área A1.

Inicialmente, para identificarmos as covariáveis meteorológicas que pudessem exercer influência nos alcances de correlação das amostras **Y**, alcance1 (**X**) e alcance2 (**Z**), executamos a função stepwise_glm que utiliza o procedimento *glm* do R para ajustar os modelos, assumindo distribuição Gama e uma função de ligação logarítmica e considerando a quadra como bloco, fixa no modelo, com nível de 5% para a inclusão e de 10% para retirada.

Nesta etapa inicial de seleção de covariáveis, talvez devido ao grande número de variáveis preditoras significativas, verificamos que a função *glm* não convergia e, neste contexto, prosseguimos executando a função stepwise_glm diminuindo o valor dos níveis de inclusão e retirada de covariáveis, além de alterarmos a função de ligação para identidade e inversa. Embora estes procedimentos tenham sido adotados e, ainda, mesmo tendo considerado um nível de entrada e retirada de 0,5%, o número de variáveis explicativas significativas continuou grande e, em determinado ponto, o algoritmo da função *glm* do R continuou a não convergir.

Para tentar resolver os problemas destacados e, tendo em vista a busca de modelos parcimoniosos, decidimos, então, executar o procedimento stepwise_glm em três etapas, nas quais assumimos uma distribuição Gama, a função de ligação logarítmica, consideramos a quadra como bloco, fixa no modelo, e níveis de entrada e retirada de covariáveis de 0,5%. Na primeira etapa, consideramos como candidatas a serem introduzidas no modelo as covariáveis medidas no dia da coleta e até sete dias atrás, ou seja, $tmax_0$, ..., $tmax_7$, $tmin_0$, ..., $tmin_7$, $tmed_0$, ..., $tmed_7$ e $pluv_0$, ..., $pluv_7$; na segunda etapa, consideramos como candidatas as covariáveis

medidas oito dias antes da coleta e até quinze dias atrás, ou seja, $tmax_8$, ..., $tmax_{15}$, $tmin_8$, ..., $tmin_{15}$, $tmed_8$, ..., $tmed_{15}$ e $pluv_8$, ..., $pluv_{15}$; na terceira etapa, consideramos as covariáveis medidas dezesseis dias antes da coleta e até vinte e um dias atrás, ou seja, $tmax_{16}$, ..., $tmax_{21}$, $tmin_{16}$, ..., $tmin_{21}$, $tmed_{16}$, ..., $tmed_{21}$ e $pluv_{16}$, ..., $pluv_{21}$. Executados estes procedimentos, conseguimos, então, encontrar três modelos.

O passo seguinte consistiu em tomarmos todas as covariáveis introduzidas nestes três modelos como variáveis explicativas candidatas para, em seguida, executarmos novamente o procedimento stepwise_glm. Desta forma, obtivemos um modelo para cada alcance de correlação, **X** e **Z**.

O modelo obtido para a variável **X** apresentou, através da tabela de análise seqüencial, ANODEV, cinco covariáveis não significativas na presença das demais e, por isso, após retirarmos a primeira variável explicativa identificada como não significativa, $tmed_0$, executamos novamente o procedimento stepwise para obtermos, então, um modelo mais parcimonioso.

Os modelos obtidos para **X** e **Z**, bem como as análises de diagnósticos, serão apresentados em **4.2.1** e **4.2.2**.

A busca de um bom modelo não se resumiu somente nos procedimentos acima citados. Executamos também a função stepwise_glm em três etapas considerando a distribuição Gama e função de ligação identidade e inversa, mas o algoritmo da função glm do R não convergiu.

4.2.1 Ajuste do modelo Alcance1_A1 - MLG

Considerando **X** uma variável aleatória com distribuição Gama e função de ligação logarítmica, obtivemos o modelo abaixo, com as respectivas estimativas dos parâmetros e seus desvios padrão:

$$\begin{aligned}\hat{\eta} = &\underset{(0,41)}{-0,29} + \ldots + \underset{(0,06)}{0,11}Q_{14}^{\bullet} + \ldots + \underset{(0,06)}{0,11}Q_{15}^{\bullet} + \underset{(0,06)}{0,11}Q_{16}^{\bullet} + \underset{(0,06)}{0,11}Q_{17}^{\bullet} + \underset{(0,06)}{0,11}Q_{22}^{\bullet} + \underset{(0,06)}{0,11}Q_{23}^{\bullet} + \underset{(0,06)}{0,11}Q_{24}^{\bullet} + \\ &+ \underset{(0,06)}{0,11}Q_{25}^{\bullet} + \underset{(0,06)}{0,11}Q_{25A}^{\bullet} + \underset{(0,06)}{0,11}Q_{26}^{\bullet} + \underset{(0,06)}{0,11}Q_{27}^{\bullet} + \underset{(0,06)}{0,24}Q_{275}^{***} + \underset{(0,06)}{0,24}Q_{277}^{***} + \underset{(0,06)}{0,24}Q_{278}^{***} + \underset{(0,06)}{0,24}Q_{279}^{***} + \\ &+ \underset{(0,06)}{0,24}Q_{287}^{***} + \underset{(0,06)}{0,11}Q_{30}^{\bullet} + \underset{(0,06)}{0,11}Q_{31}^{\bullet} + \underset{(0,06)}{0,11}Q_{33}^{\bullet} + \underset{(0,06)}{0,11}Q_{36A}^{\bullet} + \underset{(0,06)}{0,11}Q_{41}^{\bullet} + \underset{(0,06)}{0,11}Q_{42A}^{\bullet} + \underset{(0,06)}{0,11}Q_{43}^{\bullet} + \\ &+ \underset{(0,06)}{0,11}Q_{43A}^{\bullet} + \underset{(0,06)}{0,35}Q_{449}^{***} + \underset{(0,06)}{0,11}Q_{45}^{\bullet} + \underset{(0,06)}{0,11}Q_{46}^{\bullet} + \underset{(0,06)}{0,11}Q_{47A}^{\bullet} + \underset{(0,06)}{0,11}Q_{50}^{\bullet} + \underset{(0,06)}{0,11}Q_{51}^{\bullet} + \underset{(0,06)}{0,24}Q_{52}^{***} + \\ &+ \underset{(0,06)}{0,24}Q_{53}^{***} + \underset{(0,06)}{0,35}Q_{534}^{***} + \underset{(0,06)}{0,35}Q_{536}^{***} + \underset{(0,06)}{0,35}Q_{537}^{***} + \underset{(0,06)}{0,11}Q_{53A}^{\bullet} + \underset{(0,06)}{0,24}Q_{54}^{***} + \underset{(0,06)}{0,11}Q_{54A}^{\bullet} + \underset{(0,06)}{0,11}Q_{55}^{\bullet} + \\ &+ \underset{(0,06)}{0,24}Q_{55A}^{***} + \underset{(0,06)}{0,24}Q_{55B}^{***} + \underset{(0,06)}{0,24}Q_{56}^{***} + \underset{(0,06)}{0,11}Q_{62}^{\bullet} + \underset{(0,06)}{0,11}Q_{63}^{\bullet} + \underset{(0,06)}{0,24}Q_{64}^{***} + \underset{(0,06)}{0,24}Q_{64A}^{***} + \underset{(0,06)}{0,24}Q_{64B}^{***} +\end{aligned}$$

$$+\underset{(0,06)}{0,11}Q^{\bullet}_{64C}+\underset{(0,06)}{0,24}Q^{***}_{65}+\underset{(0,06)}{0,24}Q^{***}_{66}+\underset{(0,06)}{0,11}Q^{\bullet}_{66A}+\underset{(0,06)}{0,24}Q^{***}_{68}+\underset{(0,06)}{0,11}Q^{\bullet}_{69}+\underset{(0,06)}{0,11}Q^{\bullet}_{70}+\underset{(0,06)}{0,11}Q^{\bullet}_{71}+$$

$$+\underset{(0,06)}{0,11}Q^{\bullet}_{72}+\underset{(0,06)}{0,11}Q^{\bullet}_{73}+\underset{(0,06)}{0,24}Q^{***}_{74}+\underset{(0,06)}{0,11}Q^{\bullet}_{74A}+\underset{(0,06)}{0,24}Q^{***}_{75}+\underset{(0,06)}{0,11}Q^{\bullet}_{75A}+\underset{(0,06)}{0,24}Q^{***}_{79}+\ldots+\underset{(0,06)}{0,24}Q^{***}_{81}+\ldots+$$

$$+\underset{(0,06)}{0,24}Q^{***}_{83}+\ldots+\underset{(0,06)}{0,11}Q^{\bullet}_{8B}+\ldots+\underset{(0,06)}{0,11}Q^{\bullet}_{9B}-\underset{(0,0006)}{0,021}pluv^{***}_{5}-\underset{(0,0006)}{0,019}pluv^{***}_{0}-\underset{(0,0005)}{0,015}pluv^{***}_{1}+\underset{(0,004)}{0,141}t\min^{***}_{16}+$$

$$+\underset{(0,0008)}{0,002}pluv^{***}_{20}+\underset{(0,006)}{0,074}t\min^{***}_{6}+\underset{(0,016)}{0,433}t\max^{***}_{3}+\underset{(0,0007)}{0,020}pluv^{***}_{14}-\underset{(0,03)}{0,680}tmed^{***}_{3}+\underset{(0,002)}{0,080}t\max^{***}_{16}-$$

$$-\underset{(0,005)}{0,126}t\max^{***}_{7}+\underset{(0,009)}{0,207}tmed^{***}_{7}-\underset{(0,003)}{0,065}t\max^{***}_{13}+\underset{(0,007)}{0,184}t\min^{***}_{1}+\underset{(0,016)}{0,422}t\min^{***}_{3}-\underset{(0,002)}{0,066}t\max^{***}_{5}+$$

$$+\underset{(0,004)}{0,127}t\min^{***}_{21}-\underset{(0,002)}{0,0543}t\max^{***}_{21}-\underset{(0,009)}{0,263}tmed^{***}_{1}-\underset{(0,0006)}{0,012}pluv^{***}_{6}+\underset{(0,003)}{0,040}t\min^{***}_{13}+\underset{(0,029)}{0,419}tmed^{***}_{20}-$$

$$-\underset{(0,005)}{0,022}t\min^{***}_{14}+\underset{(0,004)}{0,070}t\min^{***}_{17}-\underset{(0,002)}{0,013}t\max^{***}_{4}-\underset{(0,016)}{0,252}t\min^{***}_{20}-\underset{(0,015)}{0,170}t\max^{***}_{20}+\underset{(0,0006)}{0,006}pluv^{***}_{7}+$$

$$+\underset{(0,0007)}{0,006}pluv^{***}_{21}-\underset{(0,004)}{0,041}t\max^{***}_{15}+\underset{(0,004)}{0,031}t\min^{***}_{9}-\underset{(0,008)}{0,057}tmed^{***}_{18}-\underset{(0,002)}{0,016}t\max^{***}_{9}+\underset{(0,005)}{0,027}t\max^{***}_{18}$$

onde nível: *** 0,001; ** 0,01; * 0,05; • 0,1.

Observe que as estimativas dos parâmetros regressores para as quadras significativas a um nível de 0,5% de probabilidade ou assumiram o valor 0,11, 0,24 ou 0,35, com mesmo desvio padrão 0,06.

Verificamos a adequabilidade da função de ligação logarítmica examinando a mudança ocorrida na deviance do modelo ao adicionarmos $\hat{\boldsymbol{\eta}}^2$ como covariável extra. Através da **Tabela 12**, verificamos que, a um nível de 5% de probabilidade, a função de ligação logarítmica está adequada, uma vez que a alteração da deviance não foi significativa.

Tabela 12. Verificação da adequabilidade da função de ligação logarítmica para o modelo Alcance1_A1 - MLG

Modelo	GL	Deviance Residual	Dif. de Deviances
X = fator(quadra) + pluv5 + pluv0 + pluv1 + tmin16 + pluv20 + tmin6 + tmax3 + pluv14 + tmed3 + tmax16 + tmax7 + tmed7 + tmax13 + tmin1 + tmin3 + tmax5 + tmin21 + tmax21 + tmed1 + pluv6 + tmin13 + tmed20 + tmin14 + tmin17 + tmax4 + tmin20 + tmax20 + pluv7 + pluv21 + tmax15 + tmin9 + tmed18 + tmax9 + tmax18	1466	55,258	
X = fator(quadra) + pluv5 + pluv0 + pluv1 + tmin16 + pluv20 + tmin6 + tmax3 + pluv14 + tmed3 + tmax16 + tmax7 + tmed7 + tmax13 + tmin1 + tmin3 + tmax5 + tmin21 + tmax21 + tmed1 + pluv6 + tmin13 + tmed20 + tmin14 + tmin17 + tmax4 + tmin20 + tmax20 + pluv7 + pluv21 + tmax15 + tmin9 + tmed18 + tmax9 + tmax18 + $\hat{\boldsymbol{\eta}}^2$	1465	54,460	0,798

A contribuição de cada covariável do modelo sob pesquisa está apresentada na **Tabela 13** de análise seqüencial.

Tabela 13. ANODEV Tipo I – modelo Alcance1_A1 - MLG

Fonte de variação	Deviance	GL	valor *p*
Regressão	**423,05**	**133**	**0**
factor(quadra)	0,00	99	1,00
pluv5	70,23	1	0,00
pluv0	31,91	1	5,715e-233
pluv1	30,11	1	6,377e-220
tmin16	38,87	1	2,356e-283
pluv20	25,47	1	2,374e-186
tmin6	19,33	1	6,527e-142
tmax3	24,48	1	3,522e-179
pluv14	29,43	1	5,370e-215
tmed3	10,00	1	2,365e-74
tmax16	10,10	1	4,022e-75
tmax7	18,69	1	2,629e-137
tmed7	4,98	1	6,548e-38
tmax13	13,51	1	7,856e-100
tmin1	9,73	1	2,020e-72
tmin3	1,33	1	2,788e-11
tmax5	2,92	1	6,606e-23
tmin21	4,99	1	5,102e-38
tmax21	6,74	1	1,059e-50
tmed1	27,17	1	1,041e-198
pluv6	13,25	1	6,623e-98
tmin13	6,28	1	2,192e-47
tmed20	4,10	1	1,463e-31
tmin14	2,68	1	3,585e-21
tmin17	3,71	1	1,150e-28
tmax4	2,14	1	2,989e-17
tmin20	2,03	1	2,126e-16
tmax20	1,34	1	2,345e-11
pluv7	1,63	1	1,795e-13
pluv21	0,87	1	7,755e-08
tmax15	1,00	1	7,702e-09
tmin9	1,70	1	5,091e-14
tmed18	0,92	1	3,113e-08
tmax9	0,64	1	4,144e-06
tmax18	0,77	1	4,377e-07
Resíduo	**55,258**	**1466**	
Total	**478,308**	**1599**	

A **Tabela 13** de análise seqüencial nos evidenciou através do valor *p* da regressão, a um nível de 5% de probabilidade, que o modelo se ajustou bem aos dados, havendo, ainda, a necessidade de realizarmos uma análise dos resíduos. Observamos, ainda, que todas as covariáveis são significativas na presença das demais, a um nível de 0,5% de probabilidade.

Construímos, também, os gráficos de resíduos parciais que, por não apresentarem uma tendência linear para as covariáveis $pluv_{20}$ e $pluv_{21}$, nos revelaram que as escalas destas variáveis não são satisfatórias.

A análise do gráfico de resíduos (r_i) *versus* valores ajustados, **Figura 25**, nos revelou um ajuste adequado, uma vez que os pontos estão espalhados aleatoriamente em torno de $r_i = 0$.

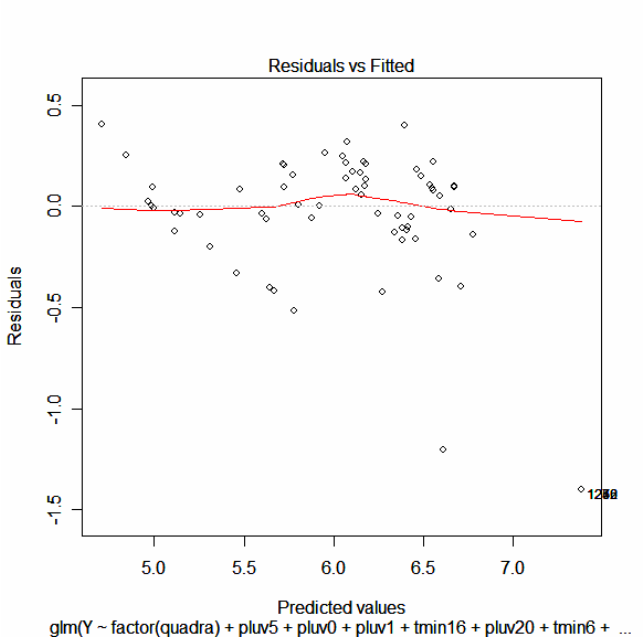


Figura 25. Gráfico dos resíduos *versus* valores preditos para o modelo Alcance1_A1 - MLG

Construímos, ainda, o gráfico normal de probabilidades para os resíduos, conforme ilustrado na, que nos confirmou um ajuste adequado do modelo adotado.

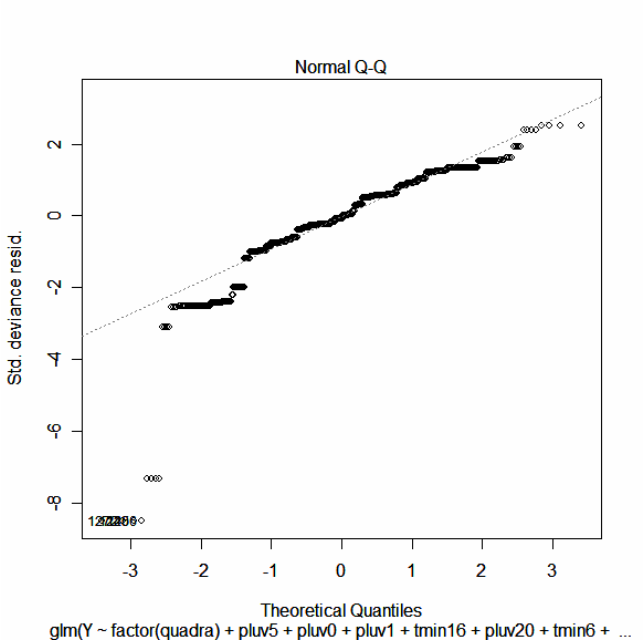


Figura 26. Gráfico normal de probabilidades para os resíduos do modelo Alcance1_A1 - MLG

Os histogramas dos resíduos da deviance e de Pearson nos revelaram uma distribuição normal assimétrica, **Figura 27**, mas, mesmo assim, vamos considerar o modelo como adequado.

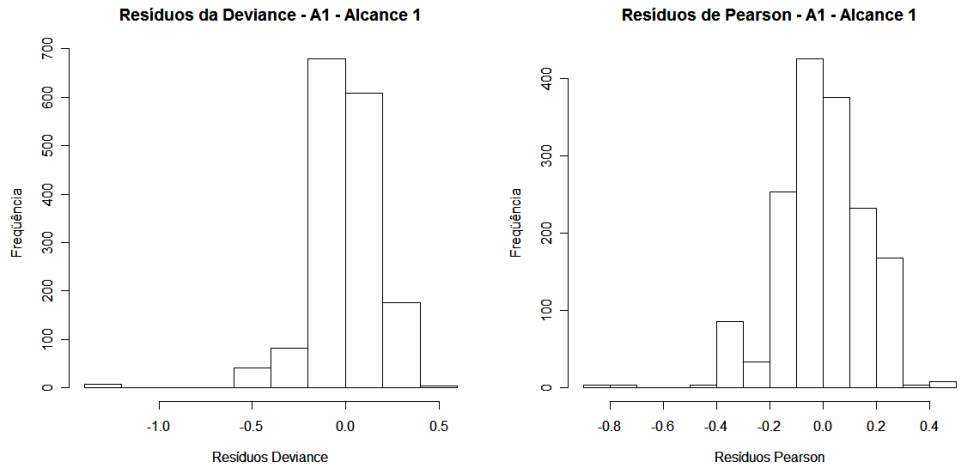


Figura 27. Histograma dos resíduos da deviance e de Pearson - modelo Alcance1_A1 - MLG

Através do gráfico da distância de Cook, uma análise informal, identificamos as observações 1240, 1256, 1272 e 1288 como outliers, **Figura 28**.

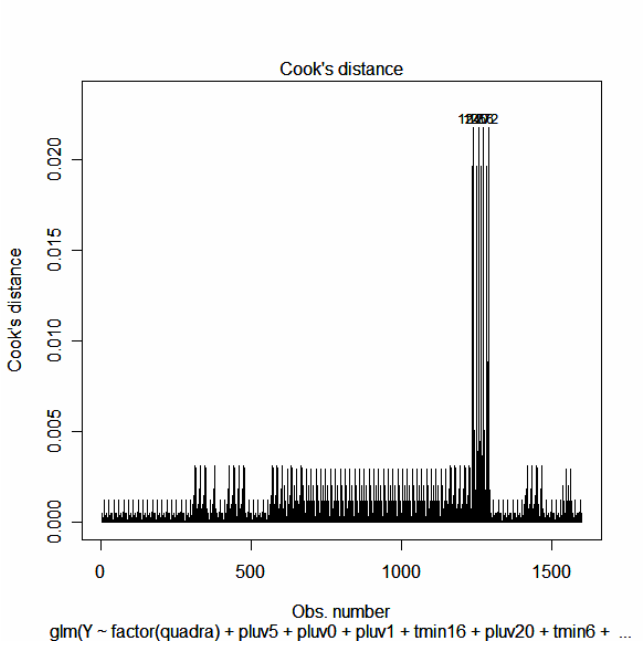


Figura 28. Gráfico da distância de Cook para o modelo Alcance1_A1 - MLG

4.2.2 Ajuste do modelo Alcance2_A1 - MLG

Supondo **Z** uma variável aleatória com distribuição Gama e função de ligação logarítmica, obtivemos como resultado o modelo abaixo especificado, com as respectivas estimativas dos parâmetros e seus desvios padrão:

$$\hat{\eta} = \underset{(0,28)}{1,25^{***}} + \ldots + \underset{(0,04)}{0,28Q_{14}^{***}} + \ldots + \underset{(0,04)}{0,28Q_{15}^{***}} + \underset{(0,04)}{0,28Q_{16}^{***}} + \underset{(0,04)}{0,28Q_{17}^{***}} + \underset{(0,04)}{0,28Q_{22}^{***}} + \underset{(0,04)}{0,28Q_{23}^{***}} + \underset{(0,04)}{0,28Q_{24}^{***}} +$$

$$+ \underset{(0,04)}{0,28Q_{25}^{***}} + \underset{(0,04)}{0,28Q_{25A}^{***}} + \underset{(0,04)}{0,28Q_{26}^{***}} + \underset{(0,04)}{0,28Q_{27}^{***}} + \underset{(0,04)}{0,46Q_{275}^{***}} + \underset{(0,04)}{0,46Q_{277}^{***}} + \underset{(0,04)}{0,46Q_{278}^{***}} + \underset{(0,04)}{0,46Q_{279}^{***}} +$$

$$+ \underset{(0,04)}{0,46Q_{287}^{***}} + \underset{(0,04)}{0,28Q_{30}^{***}} + \underset{(0,04)}{0,28Q_{31}^{***}} + \underset{(0,04)}{0,28Q_{33}^{***}} + \underset{(0,04)}{0,28Q_{36A}^{***}} + \underset{(0,04)}{0,28Q_{41}^{***}} + \underset{(0,04)}{0,28Q_{42A}^{***}} + \underset{(0,04)}{0,28Q_{43}^{***}} +$$

$$+ \underset{(0,04)}{0,28Q_{43A}^{***}} + \underset{(0,04)}{0,41Q_{449}^{***}} + \underset{(0,04)}{0,28Q_{45}^{***}} + \underset{(0,04)}{0,28Q_{46}^{***}} + \underset{(0,04)}{0,28Q_{47A}^{***}} + \underset{(0,04)}{0,28Q_{50}^{***}} + \underset{(0,04)}{0,28Q_{51}^{***}} + \underset{(0,04)}{0,46Q_{52}^{***}} +$$

$$+ \underset{(0,04)}{0,46Q_{53}^{***}} + \underset{(0,04)}{0,41Q_{534}^{***}} + \underset{(0,04)}{0,41Q_{536}^{***}} + \underset{(0,04)}{0,41Q_{537}^{***}} + \underset{(0,04)}{0,28Q_{53A}^{***}} + \underset{(0,04)}{0,46Q_{54}^{***}} + \underset{(0,04)}{0,28Q_{54A}^{***}} + \underset{(0,04)}{0,28Q_{55}^{***}} +$$

$$+ \underset{(0,04)}{0,46Q_{55A}^{***}} + \underset{(0,04)}{0,46Q_{55B}^{***}} + \underset{(0,04)}{0,46Q_{56}^{***}} + \underset{(0,04)}{0,28Q_{62}^{***}} + \underset{(0,04)}{0,28Q_{63}^{***}} + \underset{(0,04)}{0,46Q_{64}^{***}} + \underset{(0,04)}{0,46Q_{64A}^{***}} + \underset{(0,04)}{0,46Q_{64B}^{***}} +$$

$$+ \underset{(0,04)}{0,28Q_{64C}^{***}} + \underset{(0,04)}{0,46Q_{65}^{***}} + \underset{(0,04)}{0,46Q_{66}^{***}} + \underset{(0,04)}{0,28Q_{66A}^{***}} + \underset{(0,04)}{0,46Q_{68}^{***}} + \underset{(0,04)}{0,28Q_{69}^{***}} + \underset{(0,04)}{0,28Q_{70}^{***}} + \underset{(0,04)}{0,28Q_{71}^{***}} +$$

$$+ \underset{(0,04)}{0,28Q_{72}^{***}} + \underset{(0,04)}{0,28Q_{73}^{***}} + \underset{(0,04)}{0,46Q_{74}^{***}} + \underset{(0,04)}{0,28Q_{74A}^{***}} + \underset{(0,04)}{0,46Q_{75}^{***}} + \underset{(0,04)}{0,28Q_{75A}^{***}} + \underset{(0,04)}{0,46Q_{79}^{***}} + \ldots + \underset{(0,04)}{0,46Q_{81}^{***}} + \ldots +$$

$$+ \underset{(0,04)}{0,46Q_{83}^{***}} + \ldots + \underset{(0,04)}{0,28Q_{8B}^{***}} + \ldots + \underset{(0,04)}{0,28Q_{9B}^{***}} - \underset{(0,0007)}{0,004\,pluv_{5}^{***}} - \underset{(0,0005)}{0,023\,pluv_{4}^{***}} - \underset{(0,021)}{0,299t\max_{13}^{***}} +$$

$$+ \underset{(0,003)}{0,106t\min_{15}^{***}} - \underset{(0,003)}{0,082t\min_{19}^{***}} - \underset{(0,0006)}{0,005\,pluv_{2}^{***}} - \underset{(0,0006)}{0,010\,pluv_{1}^{***}} - \underset{(0,005)}{0,164tmed_{4}^{***}} + \underset{(0,0003)}{0,009\,pluv_{11}^{***}} -$$

$$- \underset{(0,0005)}{0,007\,pluv_{3}^{***}} - \underset{(0,0008)}{0,022\,pluv_{7}^{***}} - \underset{(0,002)}{0,017t\max_{7}^{***}} - \underset{(0,0008)}{0,009\,pluv_{6}^{***}} - \underset{(0,006)}{0,172t\max_{5}^{***}} - \underset{(0,012)}{0,471t\min_{0}^{***}} +$$

$$+ \underset{(0,011)}{0,393tmed_{5}^{***}} - \underset{(0,012)}{0,596t\max_{0}^{***}} + \underset{(0,024)}{1,128tmed_{0}^{***}} - \underset{(0,011)}{0,182t\max_{9}^{***}} - \underset{(0,0008)}{0,005\,pluv_{8}^{***}} + \underset{(0,023)}{0,353tmed_{9}^{***}} -$$

$$- \underset{(0,005)}{0,099t\max_{1}^{***}} - \underset{(0,012)}{0,123t\min_{9}^{***}} + \underset{(0,022)}{0,437t\max_{12}^{***}} - \underset{(0,019)}{0,184t\min_{2}^{***}} - \underset{(0,025)}{0,257t\min_{13}^{***}} + \underset{(0,0004)}{0,010\,pluv_{9}^{***}} -$$

$$- \underset{(0,004)}{0,070t\min_{1}^{***}} + \underset{(0,035)}{0,300tmed_{2}^{***}} - \underset{(0,004)}{0,097tmed_{6}^{***}} - \underset{(0,005)}{0,130t\min_{20}^{***}} - \underset{(0,013)}{0,311t\max_{21}^{***}} + \underset{(0,043)}{0,404tmed_{13}^{***}} -$$

$$- \underset{(0,015)}{0,048t\max_{2}^{***}} - \underset{(0,003)}{0,062tmed_{18}^{***}} + \underset{(0,023)}{0,467t\min_{12}^{***}} - \underset{(0,043)}{0,791tmed_{12}^{***}} + \underset{(0,028)}{0,763tmed_{21}^{***}} - \underset{(0,014)}{0,396t\min_{21}^{***}} +$$

$$+ \underset{(0,0005)}{0,008\,pluv_{21}^{***}} - \underset{(0,005)}{0,059t\max_{16}^{***}} + \underset{(0,009)}{0,064tmed_{16}^{***}} - \underset{(0,003)}{0,012t\max_{17}^{***}}$$

onde nível: *** 0,001; ** 0,01; * 0,05; • 0,1.

Observe que aqui, assim como no ajuste do alcance1, as quadras que foram significativas, a um nível de 0,5% de probabilidade, apresentaram apenas três valores para as

estimativas dos coeficientes da regressão, dados por 0,28, 0,41 e 0,46 e mesmo desvio padrão, 0,04.

Para verificarmos a adequabilidade da função de ligação acrescentamos $\hat{\boldsymbol{\eta}}^2$ como covariável extra no modelo e concluímos que, a um nível de 5% de probabilidade, a função de ligação logarítmica não está adequada, uma vez que a alteração da deviance foi significativa, conforme ilustra a **Tabela 14**.

Tabela 14. Verificando a adequabilidade da função de ligação logarítmica para o modelo Alcance2_A1 - MLG

Modelo	GL	Deviance Residual	Dif. de Deviances
Z = fator(quadra) + pluv5 + pluv4 + tmax13 + tmin15 + tmin19 + pluv2 + pluv1 + tmed4 + pluv11 + pluv3 + pluv7 + tmax7 + pluv6 + tmax5 + tmin0 + tmed5 + tmax0 + tmed0 + tmax9 + pluv8 + tmed9 + tmax1 + tmin9 + tmax12 + tmin2 + tmin13 + pluv9 + tmin1 + tmed2 + tmed6 + tmin20 + tmax21 + tmed13 + tmax2 + tmed18 + tmin12 + tmed12 + tmed21 + tmin21 + pluv21 + tmax16 + tmed16 + tmax17	**1457**	**30,5423**	
Z = fator(quadra) + pluv5 + pluv4 + tmax13 + tmin15 + tmin19 + pluv2 + pluv1 + tmed4 + pluv11 + pluv3 + pluv7 + tmax7 + pluv6 + tmax5 + tmin0 + tmed5 + tmax0 + tmed0 + tmax9 + pluv8 + tmed9 + tmax1 + tmin9 + tmax12 + tmin2 + tmin13 + pluv9 + tmin1 + tmed2 + tmed6 + tmin20 + tmax21 + tmed13 + tmax2 + tmed18 + tmin12 + tmed12 + tmed21 + tmin21 + pluv21 + tmax16 + tmed16 + tmax17 + $\hat{\boldsymbol{\eta}}^2$	**1456**	**24,7825**	**5,7598**

Este resultado nos levou a busca de novos modelos assumindo uma distribuição Gama e outras funções de ligação mas, como o algoritmo da função *glm* do R não convergiu, resolvemos prosseguir as análises de diagnósticos deste modelo, mesmo tendo identificado uma função de ligação inadequada.

A contribuição de cada covariável do modelo sob pesquisa está apresentada na **Tabela 15** de análise seqüencial.

Tabela 15. ANODEV Tipo I – modelo Alcance2_A1 - MLG

Fonte de variação	Deviance	GL	valor *p*
Regressão	**467,75**	**142**	**0**
factor(quadra)	5,684e-14	99	1,00
pluv5	72,33	1	0,00
pluv4	59,17	1	0,00
tmax13	42,14	1	0,00
tmin15	42,57	1	0,00
tmin19	3,61	1	1,668e-48
pluv2	20,03	1	4,036e-260
pluv1	15,39	1	2,069e-200
tmed4	0,25	1	1,148e-04

pluv11	10,02	1	3,264e-131
pluv3	2,96	1	4,673e-40
pluv7	9,64	1	2,668e-126
tmax7	14,49	1	8,996e-189
pluv6	6,83	1	5,520e-90
tmax5	0,02	1	0,24
tmin0	0,59	1	3,449e-09
tmed5	25,51	1	0,00
tmax0	13,48	1	1,011e-175
tmed0	4,22	1	2,324e-56
tmax9	2,41	1	6,219e-33
pluv8	16,56	1	2,145e-215
tmed9	7,69	1	4,389e-101
tmax1	1,77	1	1,343e-24
tmin9	13,73	1	5,284e-179
tmax12	2,67	1	2,942e-36
tmin2	8,11	1	1,712e-106
tmin13	0,01	1	0,49
pluv9	10,13	1	1,229e-132
tmin1	3,02	1	8,300e-41
tmed2	5,70	1	1,867e-75
tmed6	0,05	1	0,10
tmin20	3,78	1	1,300e-50
tmax21	6,92	1	3,996e-91
tmed13	2,69	1	1,754e-36
tmax2	0,01	1	0,38
tmed18	19,13	1	1,346e-248
tmin12	0,38	1	1,843e-06
tmed12	3,28	1	3,749e-44
tmed21	2,24	1	1,004e-30
tmin21	8,81	1	1,560e-115
pluv21	2,29	1	2,353e-31
tmax16	2,22	1	1,911e-30
tmed16	0,65	1	5,838e-10
tmax17	0,25	1	1,108e-04
Resíduo	**30,542**	**1457**	
Total	**498,29**	**1599**	

A **Tabela 15** de análise seqüencial nos evidenciou através do valor *p* da regressão, a um nível de 5% de probabilidade, que o modelo se ajustou bem aos dados, havendo, ainda, a necessidade de realizarmos uma análise dos resíduos.

Observe, ainda, que somente as covariáveis $tmax_5$, $tmin_{13}$, $tmed_6$ e $tmax_2$ não foram significativas na presença das demais a um nível de 0,5% de probabilidade.

Foram traçados, também, os gráficos de resíduos parciais que nos revelaram que as escalas das covariáveis $pluv_4$ e $pluv_{21}$ não são satisfatórias.

A análise do gráfico de resíduos (r_i) *versus* valores ajustados, **Figura 29**, nos revelou um ajuste inadequado, uma vez que os pontos não estão espalhados aleatoriamente em torno de $r_i = 0$, há uma pequena tendência.

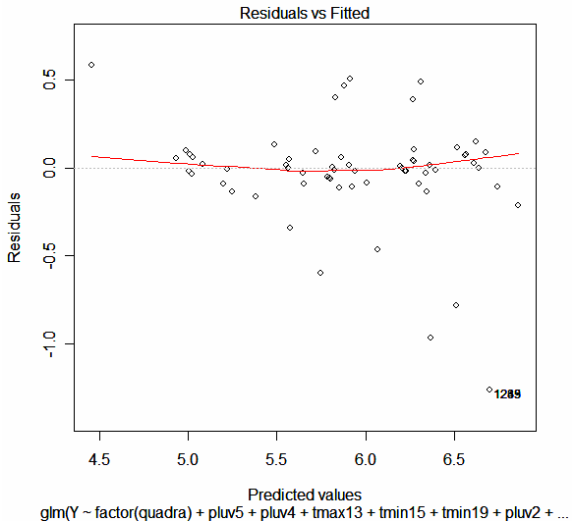


Figura 29. Gráfico dos resíduos *versus* valores preditos do modelo Alcance2_A1 - MLG

O gráfico normal de probabilidades para os resíduos, ilustrados na **Figura 30**, não apresentou a normalidade esperada. Observe, ainda, que as caudas nos revelam um ajuste inadequado do modelo aos dados.

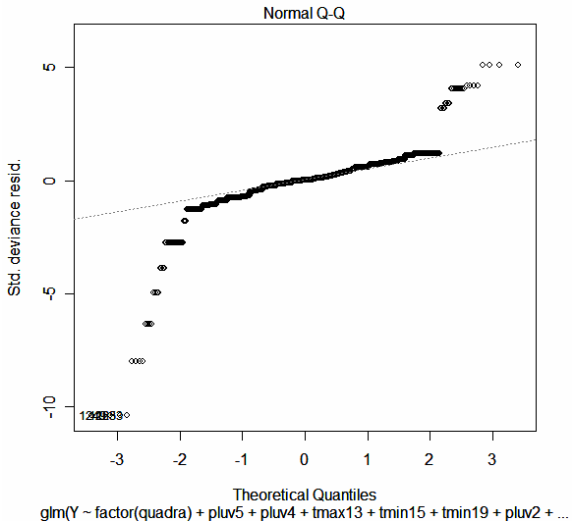


Figura 30. Gráfico normal de probabilidades para os resíduos do modelo Alcance2_A1 - MLG

Os histogramas dos resíduos da deviance e de Pearson confirmaram a não normalidade esperada, reforçando a conclusão de um ajuste inadequado, **Figura 31**.

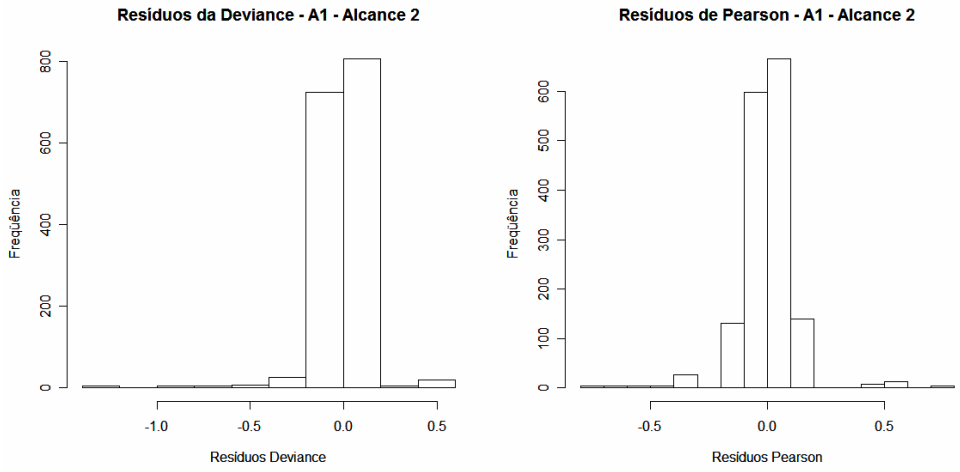


Figura 31. Histograma dos resíduos da deviance e de Pearson - modelo Alcance2_A1 - MLG

Através do gráfico da distância de Cook, **Figura 32**, verificamos que as observações 1245, 1261, 1277 e 1293 são outliers.

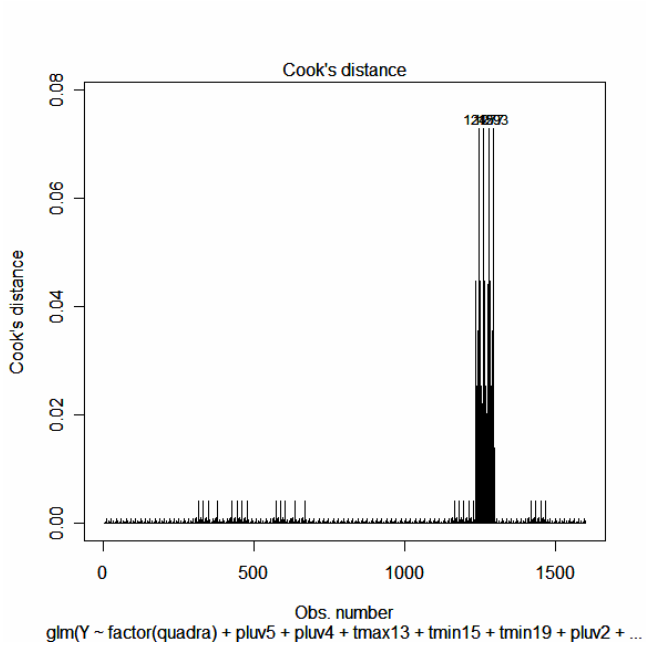


Figura 32. Gráfico da distância de Cook para o modelo Alcance2_A1 – MLG.

4.3 Principais Conclusões

As análises de diagnósticos dos ajustes dos modelos para a variável resposta **Y**, o número de fêmeas *Aedes* capturadas por armadilhas adulticidas, áreas **A1** e **A4**, via teoria de **MLG** mostraram que os modelos não se adequam bem aos dados. Foram traçados gráficos normal de probabilidades com envelope para os componentes padronizados do desvio. Notamos que os valores extremos estão distantes de uma distribuição Poisson para ambos modelos,

Mosquitrap_A1 – MLG e **Mosquitrap_A4 - MLG**. Para cada modelo ajustado foi feito, também, o teste da verificação da função de ligação e concluímos que a função de ligação logarítmica está adequada para o modelo **Mosquitrap_A1 – MLG** e, para o modelo da área **A4**, verificamos a não adequabilidade desta função de ligação. Verificamos, ainda, a existência de pontos influentes e aberrantes. Para o modelo **Mosquitrap_A1 - MLG**, detectamos os pontos: 112, 863 e 1792 como aberrantes e, também, influentes. Em seguida, excluímos os dois maiores pontos outliers deste conjunto de dados (observações 112 e 863) e ajustamos um novo modelo, cujos resultados foram semelhantes, sem alterações significativas. Neste novo ajuste, outras observações se tornaram aberrantes. Este mesmo procedimento foi executado para o modelo **Mosquitrap_A4 – MLG**, onde retiramos as observações 2014 e 2481, mas o modelo resultante também não se ajustou bem aos dados.

Para ambos modelos, **Mosquitrap_A1 – MLG** e **Mosquitrap_A4 – MLG**, construímos os gráficos de valores ajustados *versus* resíduos e observamos que o resultado foi semelhante para os dois modelos, os pontos apresentaram praticamente a mesma tendência em torno de $r_i = 0$. Os histogramas dos resíduos da deviance e de Pearson para os dois modelos nos confirmaram o ajuste inadequado dos modelos aos dados.

A busca de melhores modelos para **Y** nos levou a pesquisar a existência de uma correlação temporal para as duas áreas de análise, **A1** e **A4**. Na área **A1** detectamos uma estrutura de correlação temporal auto-regressiva de ordem 1 e, assim, prosseguimos nosso estudo efetuando uma aplicação utilizando a teoria de Equações de Estimação Generalizada para que esta correlação possa ser incorporada no modelo. Na área **A4** não aplicamos a teoria de EEG, uma vez que não identificamos a existência de correlação temporal.

Os diagnósticos realizados no modelo **Mosquitrap_A1 - GEE**, que leva em consideração a estrutura longitudinal, mais adequada a situação em questão, nos evidenciaram, também, uma qualidade de ajuste semelhante. Foram traçados histogramas dos resíduos ordinários e de Pearson, que não apresentaram a normalidade desejada. Os gráficos de valores ajustados *versus* resíduos ordinários e dos valores ajustados *versus* resíduos de Pearson não apresentaram a aleatoriedade esperada. Devido o excesso de zeros, o modelo em questão ajustou valores entre 0 até 1.5, quando temos observados valores de 0 até 8, implicando em grandes valores para os resíduos. Fato este que, talvez, justifique o comportamento anômalo dos resíduos.

Ainda analisando os três modelos ajustados, verificamos que as covariáveis $pluv_{18}$ e $tmax_9$ estão presentes nos modelos **Mosquitrap_A1 – MLG** e **Mosquitrap_A1 – EEG** com coeficiente, desvio padrão e nível de significância bem próximos. A covariável $pluv_{13}$ está presente nos modelos **Mosquitrap_A1 – MLG** e **Mosquitrap_A4 – MLG** com coeficiente,

desvio padrão e nível de significância bem próximos. As demais variáveis explicativas introduzidas são distintas.

Para confirmar os resultados obtidos, extraímos uma amostra aleatória de tamanho 10 e ajustamos três novos modelos: **Mosquitrap_A1_10 – MLG, Mosquitrap_A1_10 – EEG** e **Mosquitrap_A4_10 - MLG** . O modelo **Mosquitrap_A1_10 – MLG** apresentou resultados diferentes aos do obtido no **Mosquitrap_A1 – MLG**. O modelo **Mosquitrap_A1_10 – EEG** também apresentou resultados diferentes aos do obtido no modelo **Mosquitrap_A1 – EEG**, apenas uma covariável, $tmax_9$, foi introduzida em ambos modelos mas, com coeficiente, desvio padrão e nível de significância diferentes, as demais variáveis explicativas foram todas diferentes. O modelo **Mosquitrap_A4_10 – MLG** também apresentou resultados diferentes do modelo **Mosquitrap_A4 – MLG**, apenas a covariável $tmin_7$ foi introduzida em ambos modelos mas, com coeficiente, desvio padrão e nível de significância diferentes, as demais variáveis explicativas foram todas diferentes. Estes fatos nos indicam a existência de dados com grande dispersão para as duas áreas de análises.

A busca de melhores modelos utilizando-se as teorias estatísticas clássicas, MLG e EEG, prosseguiram, mais sem sucesso.

Neste contexto, utilizamos a Geoestatística para analisar a variação espacial do *Aedes*, mais especificamente, efetuamos um estudo da estrutura de dependência espacial deste fenômeno para que ações de controle e de prevenção mais precisas sejam efetuadas.

Através da análise Geoestatística conseguimos identificar que a direção privilegiada de variabilidade do fenômeno em estudo para a área **A1** é Noroeste-Sudeste para todas as semanas da análise e, para a área **A4**, a tendência é Norte-Sul, exceto para as semanas 6, 14, 15 e 21, onde a direção de maior continuidade espacial é concordante com a área **A1**. Entretanto, é importante observar que estes resultados podem estar sendo afetados pela amostragem efetuada em cada área. Embora tenhamos identificado uma direção privilegiada de variabilidade para todas as semanas de análise nas duas áreas de amostragem, não conseguimos construir semivariogramas adequados para as duas direções para a maior parte das semanas, tanto para a área **A1** como **A4**, nos forçando a assumir, então, um modelo isotrópico para estas semanas em questão. A desvantagem de assumirmos um fenômeno isotrópico ou invés de anisotrópico será refletida na interpolação/previsão, ou seja, na inferência sobre a realização do processo em localizações não medidas, gerando estimativas pobres, uma vez que não incorporamos no modelo adotado a direção privilegiada de variabilidade.

Esta análise, nos revelou, ainda, raios de correlação de amostras variando de 146 a 868 metros para a área **A1** e de 27 até 280 metros para a área **A4**, resultados importantes a serem considerados e analisados para que ações de controle e de prevenção mais seguras e corretas sejam efetuadas. Vale ressaltar que o maior valor do alcance de correlação das amostras, 868 metros, pode ser concordante com os resultados divulgados no artigo de NILDIMAR et al. (2003), onde foi identificado que o alcance de vôo do *Aedes* pode atingir até 800 metros. Observe que, no nosso caso, temos o alcance de correlação das amostras e no artigo, temos o alcance de vôo do *Aedes*. A relação provável existente entre estes resultados é que, se for capturada uma fêmea em determinado ponto amostral e se neste local existirem ovos, a partir do nascimento dos mosquitos, alguns podem permanecer próximos ao local de nascimento e outros podem voar até 868 metros, alcance de correlação das amostras, assim, o resultado seria condizente com o artigo de NILDIMAR et al. (2003). Caso contrário, podemos concluir que os alcances de correlação obtidos são devidos a locais próximos favoráveis ao depósito de ovos ou que fornecem alimento às fêmeas, existindo, então, uma quantidade razoável de fêmeas sobrevoando esta área de correlação.

Os valores dos alcances de correlação obtidos nos permitem concluir, ainda, que podemos aumentar o espaçamento do local de instalação das armadilhas em novas pesquisas/estudos, contribuindo na diminuição dos custos utilizados.

Através das análises dos resíduos verificamos que a maior parte dos modelos ajustados via teoria de Geoestatística representam adequadamente a variabilidade espacial do fenômeno em estudo, exceto os modelos das semanas 9, 11, 12, 13 e 19, área **A1, e** semanas 7, 13, 20 e 21, área **A4**, que geraram valores amostrais superestimados ou subestimados devido ao excesso de zeros existentes nas duas áreas de amostragem, ou, de amostragens inadequadas ou com poucas observações, ou, ainda, do modelo adotado que pode não estar representando de forma adequada a variabilidade espacial da variável em estudo.

Embora este fato tenha ocorrido para algumas semanas, concluímos que a utilização da Geoestatística é mais adequada para análise do número de fêmeas *Aedes* capturadas, pois, além da maior parte dos modelos estarem bem ajustados, conseguimos identificar o raio de correlação das amostras e a direção privilegiada de variabilidade deste fenômeno, dados importantes não revelados quando utilizamos as teorias da estatística clássica para modelagem.

Vale ressaltar todos os resultados são provenientes da escolha do semivariograma considerado supostamente o “mais apropriado”. Na Geoestatística não existe uma receita para se produzir semivariogramas adequados aos modelos ideais. A construção do semivariograma é pessoal e subjetiva, e, ainda, exige uma certa experiência por parte do pesquisador. Neste

contexto, outras análises devem gerar resultados diferentes, valendo-se da prática/conhecimento de outros pesquisadores.

A identificação dos raios de maior e menor espalhamento do fenômeno em estudo nos levou a procura de modelos que pudessem explicar a relação entre os valores obtidos para estes alcances com as variáveis meteorológicas via MLG, uma vez que não identificamos a existência uma dependência temporal em ambas as áreas de amostragem.

As análises de diagnósticos realizadas para o modelo da variável **X**, alcance de maior espalhamento, área **A1**, nos evidenciaram um ajuste adequado, com um grande número de variáveis meteorológicas influenciando neste alcance de correlação. Realizamos o teste da adequacidade do uso da função de ligação logarítmica que resultou em valores favoráveis à sua utilização. Foram traçados os gráficos dos valores observados *versus* valores ajustados que nos mostraram a aleatoriedade dos pontos em torno de $r_i = 0$. Os resultados foram confirmados, ainda, com o gráfico normal de probabilidades que nos evidenciou uma nuvem de pontos bem próximas da reta 45º, o desejado para um modelo adequado.

Entretanto, as análises de diagnósticos do ajuste do modelo para a variável resposta **Z**, o alcance de menor espalhamento, área **A1**, mostraram que o modelo ajustado não está adequado. Realizamos o teste da função de ligação que nos evidenciou que a função de ligação logarítmica não é adequada. Neste sentido, tentamos ajustar novos modelos, com outras funções de ligação, mas tivemos problemas de convergência com a função *glm* do R, fato que nos conduz a efetuarmos novos estudos para que seja possível a aplicação de outras metodologias, como por exemplo o uso de modelos assimétricos, ou, ainda, no desenvolvimento de um novo algoritmo de estimação.

Para a área **A4** não conseguimos encontrar um modelo para os alcances, pois encontramos problemas de convergência com a função *glm* do R o que nos leva ao estudo do uso de novas metodologias, métodos mais robustos de estimação ou, ainda, no desenvolvimento de novos algoritmos de procedimentos iterativos de estimação.

5 – PROPOSTAS FUTURAS

As análises de diagnósticos dos ajustes realizados via teoria de MLG e de EEG nos revelaram que os modelos elaborados não se adequaram bem aos dados, já que apresentaram grandes resíduos. Além disso, verificamos também a existência de dados com grande dispersão através de novos ajustes realizados em uma amostra extraída do banco de dados. Estes fatos trazem como principal conseqüência modelos com má qualidade preditiva.

Neste contexto, utilizamos a teoria de Geoestatística que, além de ajustar um grande número de modelos representando adequadamente a variabilidade espacial do número de fêmeas *Aedes* capturadas, nos permitiu o conhecimento de dados importantes para que ações de controle mais precisas sejam efetuadas. Estes fatos nos permitem concluir que modelos ajustados via Geoestatística para nossa principal variável em estudo é mais adequado, uma vez que nos fornece estimativas mais precisas e conseqüentemente, modelos com melhores capacidades preditivas.

Com o objetivo de explicitar mais detalhadamente o comportamento do *Aedes*, propomos a continuidade do nosso estudo conforme especificado abaixo:

- realização do procedimento de krigagem, ou seja, interpolação do processo em localizações não medidas;
- construção de mapas oriundos do processo de krigagem para a visualização das áreas de risco;
- estudo e aplicação das técnicas da Geoestatística Multivariada, buscando relacionar espacialmente o número de fêmeas *Aedes* capturadas com as variáveis meteorológicas;
- realização da cokrigagem a partir dos modelos ajustados via Geoestatística Multivariada e a subseqüente construção dos mapas para a identificação das áreas de risco;
- comparação dos resultados da krigagem e da cokrigagem;
- estudo e aplicação da krigagem e da cokrigagem via simulação;
- estudo e aplicação das técnicas da análise espaço-temporal;
- realização de uma regressão polinomial para tentar relacionar o número de fêmeas *Aedes* capturadas com as variáveis meteorológicas;
- elaboração de novos modelos para explicar a relação entre os alcances obtidos com as variáveis meteorológicas.

6 - REFERÊNCIAS

BARATA, E.A.M.F.; COSTA, A.I.P.C.; CHIARAVALLOTI-NETO, F.; GLASSER, C.M.; BARATA, J.M.S.; **NATAL** D. **População de Aedes aegypti (l.) em área endêmica de dengue, Sudeste do Brasil**. Rev Saúde Pública, 2001. 35(3):237-242.

BARRETO, M.C.M. **Modelos Lineares Generalizados** – Notas de Aula. Universidade Federal de São Carlos, Departamento de Estatística, 2005.

CAMARGO, E.C.G; FELGUEIRAS, C.A.; MONTEIRO, A.M.V. **A Importância da Modelagem da Anisotropia na Distribuição Espacial de Variáveis Ambientais Utilizando Procedimentos Geoestatísticos**. Anais X SBSR, Foz do Iguaçu, INPE – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brasil, 2001. p. 395-402.

CAMARGO, E.C.G. **Desenvolvimento, Implementação e Teste de Procedimentos Geoestatísticos (Krigeagem) no Sistema de Processamento de Informações Georreferenciadas (Spring).** Dissertação (Mestrado). INPE – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brasil, 1997, 146p.

CHIARAVALLOTI-NETO, F.; DIBO, M.R.; EIRAS, A.E.; FAVARO, E.A.; MOURA, M.S.A.; **Estudo da relação entre indicadores entomológicos para *Aedes (Stegomyia) aegypti* obtidos de armadilhas adulticidas, de oviposição e de coleta de adultos, em área da região noroeste do estado de São Paulo.** Projeto de Pesquisa. Faculdade de Medicina de São José do Rio Preto, Departamento de Epidemiologia e Saúde Coletiva, São José do Rio Preto, SP, 2004.

DEMÉTRIO, C.G.B. **Modelos Lineares Generalizados em Experimentação Agronômica**. ESALQ – USP – Piracicaba, SP, 2002.

DIGGLE, P.J.; RIBEIRO JUNIOR, P.J. **Model-based Geoestatistics**. Departamento de Estatística, Universidade Federal do Paraná, Brasil. 14º SINAPE – Caxambu, 2000.

EIRAS, A.E. Armadilha para captura de mosquitos. 2002. Patente: Privilégio e Inovação. nº PI0203907-9, Armadilha para captura de mosquitos. 05 de set de 2002 (Depósito).

HORTON, N.J.; LIPSITZ, S. **Review of Software to Fit Generalized Estimating Equation Regression Model**. The American Statistical Association, 1999, volume 53, p. 160-169.

HUIJBREGTS, C.J. **Regionalized Variables and Quantitative Analysis of Spatial Data**. In: Davis, J.C. & McCullagh, M.J. (ed) Display and analysis of spatial data. New York, John Wiley, 1975. p.38-53.

JIAN, X.; OLEA, R.A.; YU, Y.S. **Semivariogram Modeling by Weighted Least Squares**. Computers & Geosciences, volume 22, nº 4, 1996, p.387-397.

JOHNSTON, G. **Repeated Measures Analysis with discrete Data Using the SAS System**. SAS Institute Inc., Cary, NC, from SUGI Proceedings, 1996.

LANDIM, P.M.B. **Análise Estatística de dados geológicos**. São Paulo: ed. Unesp. 2ª ed, 2003.

MARC, S; TOBIAS, A.; MUNOZ, P.; CAMPBELL, M.J. **A Gee Moving Average Analysis of the Relationship between Air Pollution and Mortality for Asthma in Barcelona, Spain**. Statistics in Medicine, n. 18, 1999. p. 2077-2986.

MELLO, J.M.; BATISTA, J.L.F.; RIBEIRO JUNIOR, P.J.; OLIVEIRA, M.S. **Ajuste e seleção de modelos espaciais de semivariograma visando a estimativa volumétrica de Eucalyptus grandis**. Scientia Forestalis, n. 69, dez. 2005, p. 25-37.

MAINDONALD, J.H. **Using R for Data Analysis and Graphics Introduction**, **Code and Commentar**. Centre for Bioinformation Science, Australian National University, 2004.

NETO, F.C. **Estudo da relação entre indicadores entomológicos para *Aedes* (*Stegomyia*) *aegypti* obtidos de armadilhas adulticidas, de oviposição e de coleta de adultos, em área da região noroeste do estado de São Paulo.** Projeto de Pesquisa. Departamento de Epidemiologia e Saúde Coletiva da Faculdade de Medicina de São José do Rio Preto, 2004.

NILDEMAR, A.H.; SILVA, W.C.; LEITE, P.J.; GONÇALVES, J.M.; LOUNIBOS, L.P.; OLIVEIRA, R.L. **Dispersal of *Aedes aegypti* and *Aedes albopictus* (Diptera: Culicidae) in Urban Endemic Dengue Area in the State of Rio de Janeiro, Brazil.** Mem. Inst. Oswaldo Cruz, Rio de Janeiro, vol. 98(2), 2003, p. 191-198.

OLEA, R.A. **Fundamentals of semivariograms estimation, modeling and usage**. In: Yarus, J.M. & Chambers, R.L. (eds) Stochastic modeling and geostatistics - principles, methods and case studies. Tulsa, AAPG (AAPG Computer Application in Geology no 3), 1994, p. 27-36.

OLIVEIRA, G. M.; BARRETO, M.C.M. **Uma análise de dados categorizados longitudinais de um programa de atividade física na qualidade de vida de mulheres com osteoporose**. Revista Matemática e Estatística, São Paulo, v. 21, n.2, 2003. p. 43-54.

PAULA, G.A. **Modelos de Regressão com apoio computacional**. Instituto de Matemática e Estatística – USP, 2004.

SOARES, A. **Geoestatística para as Ciências da Terra e do Ambiente**. Instituto Superior Técnico, ed. IST Press, Lisboa, Portugal, 2006.

APÊNDICE A

Alguns resultados de Modelos Lineares Generalizados

A1) Função escore

$$U(\beta_j) = \sum_{i=1}^{n} \frac{\partial \ell(\theta_i; y_i, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} =$$

$$= \sum_{i=1}^{n} \left\{ \frac{d\ell_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} \right\} =$$

$$= \sum_{i=1}^{n} \frac{1}{a(\phi)} [y_i - b'(\theta_i)] \frac{1}{\frac{d\mu_i}{d\theta_i}} \frac{d\mu_i}{d\eta_i} x_{ij} =$$

$$= \sum_{i=1}^{n} \frac{1}{a(\phi)} [y_i - \mu_i] \frac{1}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} =$$

$$= \sum_{i=1}^{n} \frac{w_i}{\phi} [y_i - \mu_i] \frac{1}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} =$$

$$= \sum_{i=1}^{n} \frac{1}{\phi} x_{ij} W_i \frac{d\eta_i}{d\mu_i} [y_i - \mu_i]$$

A2) Matriz de Informação esperada de Fisher

$$\Im(\beta)_{jk} = E\left(\frac{\partial \ell^2(\theta; y, \phi)}{\partial \beta_j \partial \beta_k} \right) = E(U_j U_k) =$$

$$= \sum_{i=1}^{n} \frac{1}{[a_i(\phi)]^2} E(Y_i - \mu_i)^2 \frac{1}{[V(\mu_i)]^2} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{ij} x_{ik} =$$

$$= \sum_{i=1}^{n} \frac{1}{[a_i(\phi)]^2} a_i(\phi) V(\mu_i) \frac{1}{[V(\mu_i)]^2} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{ij} x_{ik} =$$

$$= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{ij} x_{ik} =$$

$$= \sum_{i=1}^{n} \frac{1}{\phi} x_{ij} W_i x_{ik} =$$

APÊNDICE B

Alguns resultados de Geoestatística

B1) Função covariância:

$$
\begin{aligned}
C(u) &= \rho(u)\sqrt{Var[Y(x)]Var[Y(x')]} = \\
&= \rho(u)\sqrt{\{E[Y^2(x)]-\mu^2\}\{E[Y^2(x')]-\mu^2\}} = \\
&= \rho(u)\sqrt{E[Y^2(x)]^2 - E[Y^2(x)]\mu^2 - E[Y^2(x)]\mu^2 + \mu^4} = \\
&= \rho(u)\sqrt{E[Y^2(x)]^2 - 2E[Y^2(x)]\mu^2 + \mu^4} = \\
&= \rho(u)\sqrt{\{E[Y^2(x)]-\mu^2\}^2} = \\
&= \rho(u)\left\{\{E[Y^2(x)]-\mu^2\}^2\right\}^{1/2} = \\
&= \rho(u)\{E[Y^2(x)]-\mu^2\} = \\
&= \rho(u)Var[Y(x)] = \rho(u)\sigma^2.
\end{aligned}
$$

B2) A função semivariograma $\gamma(u)$ expressa em termos da variância σ^2 e da função covariância C(x, x') = C(x, $x+u$) = C(u):

$$
\begin{aligned}
\gamma(u) &= \frac{1}{2}E\{[Y(x+u)-Y(x)]^2\} = \\
&= \frac{1}{2}\{E\{Y^2(x+u)-2Y(x+u)Y(x)+Y^2(x)\}\} = \\
&= \frac{1}{2}\{E\{Y^2(x+u)\}-2E\{Y(x+u)Y(x)\}+E\{Y^2(x)\}\},
\end{aligned}
$$

mas, admitindo-se a estacionariedade, temos que E{$Y^2(x+u)$} = E{$Y^2(x)$} e, ainda, como a função covariância é expressa por:

$$C(Y(x+u), Y(x)) = C(u) = E\{Y(x+u)\,Y(x)\} - E\{Y(x+u)\}\,E\{Y(x)\} =$$

$$= E\{Y(x+u)\,Y(x)\} - \mu^2.$$

Este resultado implica em $E\{Y(x+u)\}\,E\{Y(x)\} = C(u) + \mu^2$.

Ainda, sabemos que $Var(Y(x)) = E\{Y^2(x)\} - \mu^2$. Assim, $E\{Y^2(x)\} = Var(Y(x)) + \mu^2 = \sigma^2 + \mu^2$.

Com estes resultados, podemos reescrever a função semivariograma como segue:

$$\gamma(u) = \frac{1}{2}\left\{E\left\{Y^2(x+u)\right\} - 2E\left\{Y(x+u)Y(x)\right\} + E\left\{Y^2(x)\right\}\right\}$$
$$= \frac{1}{2}\left\{\sigma^2 + \mu^2 - 2\left[C(u) + \mu^2\right] + \sigma^2 + \mu^2\right\} =$$
$$= \frac{1}{2}\left\{2\sigma^2 + 2\mu^2 - 2C(u) - 2\mu^2\right\} =$$
$$= \frac{1}{2}\left\{2\sigma^2 - 2C(u)\right\} =$$
$$= \sigma^2 - C(u) =$$
$$= \sigma^2 - \sigma^2\rho(u) =$$
$$= \sigma^2\left\{1 - \rho(u)\right\}.$$

Desta forma, mostramos que para um processo com estrutura de covariância estacionária, o variograma se reduz a função semivariograma $\gamma(u) = \frac{1}{2}\left\{2\sigma^2 - 2C(u)\right\} = \sigma^2 - C(u)$, que é a metade da *função* variograma.

APÊNDICE C

Funções stepwise desenvolvidas em *R*

C1) Função stwpwise_glm

```
stepwise_glm = function(Y, ve, pdados, familia, nivel, nivel_ret, termo_obr, num_par_fixos)
{
ve_i  = c()
ve_i  = termo_obr
linha = num_par_fixos
linhav_e = 1 + linha
linha_inicial = 1 + linha
controle = "próximo"
repeat
{
p_valor  = c()
for (i in 1:length(ve))
 {
     ven        = c() ## vetor com as novas variaveis explicativas
     ven        = c(ve_i, ve[i])
     cat(i,"º Ajuste: y = ",paste(ven, collapse= "+"),"\n")
     formula    = as.formula(paste("Y ~ ", paste(ven, collapse= "+")))
     modelo     = glm(formula, family=familia, data=pdados)
     p_valor    = c(p_valor,coef(summary(modelo))[linhav_e,4]) ## vetor com os p-valores
  }

 cat("\n","\n");cat("Dado que o modelo tem as variáveis: ","\n","\n",paste("Y ~ ", paste(ve_i, collapse=
   "+")),"\n","\n");cat("O p-valor das variáveis candidatas é: ","\n","\n")
 for (j in 1:length(ve))
   {
    ##cat(ve[j]," ",p_valor[j],"\n")
     cat(ve[order(p_valor)[j]]," ",p_valor[order(p_valor)[j]],"\n")
   }

if (p_valor[order(p_valor)[1]] <= nivel)                                    {ve_i = c(ve_i, ve[order(p_valor)[1]]);
    cat("\n","A variável mais significativa foi: ",ve[order(p_valor)[1]],"\n");        ve=ve[-order(p_valor)[1]];
    cat("\n","VARIÁVEL(IS) ADICIONADA(S): ",ve_i,"\n","\n")}                                  else {
    cat("\n","PARAR !!!","\n"); controle="parar"}
if (controle == "parar") {break}

formula = as.formula(paste("Y ~ ", paste(ve_i, collapse= "+")))
modelo  = glm(formula, family=familia, data=pdados)

if (length(coef(modelo)) > linha_inicial) ## pois vou verificar a partir da 2ª v.e. incluída
{
 verificar = "S"
 while (verificar == "S")
  {
    v_e_retirada = c()
    indice = -1
    verificar = "N"

    for (z in linha_inicial:(length(coef(modelo))))
      {
        if (coef(summary(modelo))[z,4] >= nivel_ret)
          {
            v_e_retirada = coef(modelo)[z]
            indice = 0
```

```
            verificar = "S"
            cat("Por se tornar não significativa, será retirada: ",names(v_e_retirada),"\n")
           }
      }

   if (indice == 0) ## há variáveis não significativas
      {
         for (z in 1:length(ve_i))
           {
             if (ve_i[z]==names(v_e_retirada))
               {
                  indice=z  ### índice do vetor ve_i que contém a var. que será retirada
                  cat("Índice da v.retirada = ",indice,"\n")
                  z=length(ve_i)  ### finalizando o for
                }
           }
         ve_i = ve_i[-indice]
         formula = as.formula(paste("Y ~ ", paste(ve_i, collapse= "+")))
         modelo  = glm(formula, family=familia, data=pdados)
      }
  }
}
linhav_e = length(coef(modelo))+1 ## linha que estará o p-valor da próxima var. explicativa
if (length(ve)==0) {break} ## se o vetor de variáveis explicativas estiver vazio, devo finalizar a função
}
formulaObtida   = paste("Y ~ ", paste(ve_i, collapse= "+"))
return(cat("\n","Fórmula obtida = ", formulaObtida,"\n"))
}
```

C2) Função stwpwise_gee

```
stepwise_gee = function(Y, ve, pid, pdados, familia, estrutura, nivel, nivel_ret, termo_obr, num_par_fixos)
{
library(geepack)
ve_i  = c()
ve_i  = termo_obr
linha = num_par_fixos
linhav_e = 1 + linha
linha_inicial = 1 + linha
controle = "próximo"
repeat
{
p_valor  = c()
for (i in 1:length(ve))
  {
      ven        = c()
      ven        = c(ve_i, ve[i])
      cat(i,"º Ajuste: y = ",paste(ven, collapse= "+"),"\n")
      formula    = as.formula(paste("Y ~ ", paste(ven, collapse= "+")))
      modelo_geeP = geeglm(formula, id=pid, data=pdados, family = familia, corstr=estrutura)
      p_valor    = c(p_valor,coef(summary(modelo_geeP))[linhav_e,4]) ## vetor com os p-valores
   }

 cat("\n","\n");cat("Dado que o modelo tem as variáveis: ","\n","\n",paste("Y ~ ", paste(ve_i, collapse=
    "+")),"\n","\n");cat("O p-valor das variáveis candidatas é: ","\n","\n")
 for (j in 1:length(ve))
    {
      cat(ve[order(p_valor)[j]]," ",p_valor[order(p_valor)[j]],"\n")
    }
```

```
if (p_valor[order(p_valor)[1]] <= nivel)                                    {ve_i = c(ve_i, ve[order(p_valor)[1]]);
     cat("\n","A variável mais significativa foi: ",ve[order(p_valor)[1]],"\n");          ve=ve[-order(p_valor)[1]];
     cat("\n","VARIÁVEL(IS) ADICIONADA(S): ",ve_i,"\n","\n")}                                          else {
     cat("\n","PARAR !!!","\n"); controle="parar"}

if (controle == "parar") {break}
formula = as.formula(paste("Y ~ ", paste(ve_i, collapse= "+")))
modelo_geeP = geeglm(formula, id=pid, data=pdados, family = familia, corstr=estrutura)
if (length(coef(modelo_geeP)) > linha_inicial) ## pois vou verificar a partir da 2ª v.e. incluída
{
 verificar = "S"
 while (verificar == "S")
  {
    v_e_retirada = c()
    indice = -1
    verificar = "N"
    for (z in linha_inicial:(length(coef(modelo_geeP))))
      {
        if (coef(summary(modelo_geeP))[z,4] >= nivel_ret)
          {
            v_e_retirada = coef(modelo_geeP)[z]
            indice = 0
            verificar = "S"
            cat("Por se tornar não significativa, será retirada: ",names(v_e_retirada),"\n")
          }
      }

    if (indice == 0)
       {
          for (z in 1:length(ve_i))
            {
              if (ve_i[z]==names(v_e_retirada))
                {
                   indice=z
                   cat("Índice da v.retirada = ",indice,"\n")
                   z=length(ve_i)
                 }
            }
          ve_i = ve_i[-indice]
          formula = as.formula(paste("Y ~ ", paste(ve_i, collapse= "+")))
          modelo_geeP = geeglm(formula, id=pid, data=pdados, family=familia, corstr=estrutura)
       }
  }
}
linhav_e = length(coef(modelo_geeP))+1 ## linha que estará o p-valor da próxima var. explicativa
if (length(ve)==0) {break} ## se o vetor de variáveis explicativas estiver vazio, devo finalizar a função
}
formulaObtida   = paste("Y ~ ", paste(ve_i, collapse= "+"))
return(cat("\n","Fórmula obtida = ", formulaObtida,"\n"))
}
```