Universidade Federal do Rio de Janeiro
Centro de Ciências Matemáticas e da Natureza
Instituto de Matemática
Departamento de Métodos Estatísticos

# Aplicação de técnicas de Estatística Espacial na modelagem da relação existente entre temperatura e umidade relativa do ar.

Caroline Teixeira de Castro

Priscila da Fonseca Leitão

PROJETO FINAL DE CURSO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO TÍTULO DE ATUÁRIO/ESTATÍSTICO.

Rio de Janeiro

2013

## Sumário

## Índice de Figuras

**Resumo**

Este trabalho aborda o estudo de técnicas de Estatística Espacial em comparação à Análise de Regressão para identificação de um modelo que melhor descreva a relação existente entre temperatura e umidade relativa do ar. As variáveis de interesse estão indexadas no espaço e, portanto, visamos modelar uma possível dependência espacial, que não seria tratada no modelo de Regressão usual. Iremos considerar um modelo de geoestatística bivariado que será definido de forma condicional Y e Y|X. Neste trabalho, utilizaremos inferência Bayesiana para inferir sobre parâmetros dos modelos e o pacote WinBUGS para implementação.

# 1 Introdução

O estudo da relação existente entre temperatura e umidade relativa do ar é fundamental para viabilidade de técnicas aplicadas a diversas áreas de pesquisa. No cultivo de alimentos em geral, como o milho e a soja, na produção animal, como a viabilidade embrionária durante a incubação de aves, na promoção da qualidade de vida, como o desempenho térmico de habitações, entre outras áreas de pesquisa, a relação existente entre temperatura e umidade relativa do ar está presente como fator determinístico.

Neste projeto, estudaram-se os valores da temperatura, em graus Celsius, e da umidade relativa do ar de 38 estações no Ceará, medidas em 26/09/2008 às 18 horas. Os dados são provenientes do INPE (Instituto Nacional de Pesquisas Espaciais), cujas atividades buscam demonstrar que a utilização da ciência e da tecnologia espacial pode influir na qualidade de vida da população brasileira e no desenvolvimento do país.

Este estudo teve como objetivo a proposição de um modelo que, comparado à Regressão Linear, melhor descreva a relação existente entre temperatura e umidade relativa do ar. Um modelo que não considere apenas uma relação funcional simples entre as variáveis, mas que pondere as características espaciais dos dados, avaliando a sua localização.

A Estatística Espacial é dividida de acordo com os tipos de observações associadas ao espaço em que elas são observadas. De uma forma geral, a Estatística Espacial contém três grandes áreas, descritas no capítulo 4, a saber: geoestatística, dados de área e processos pontuais. Neste projeto, trabalharemos com dados geoestatísticos.

No próximo capítulo, apresentamos o modelo de Regressão Linear, que não acomoda dependência espacial. No capítulo 3, definimos Processos Estocásticos e, particularmente, os Processos Gaussianos. No capítulo 4, descrevemos o modelo de Estatística Espacial, em que incorporamos ao modelo a localização dos dados. No capítulo 5, apresentamos a aplicação aos dados de ambos os modelos citados, concluindo, no último capítulo, que o modelo espacial é aquele, entre os modelos analisados, que melhor descreve a relação existente entre temperatura e umidade relativa do ar.

# 2 Análise de Regressão

Considere um sistema em que as quantidades variáveis sofram mudança. Neste caso, é de interesse examinar a relação de covariáveis com uma variável resposta de interesse.

Pode de fato existir uma relação funcional simples entre variáveis. Mas, na maioria dos processos, esta é a exceção e não a regra. Muitas vezes existe uma relação funcional que é muito complicada de entender ou de descrever em termos simples.

Nesse caso, podemos desejar aproximar essa relação funcional por alguma função matemática simples, tal como um polinômio, que contenha as variáveis apropriadas e que se aproxima da verdadeira função, utilizando algumas das variáveis envolvidas.

Ao examinar tal função somos capazes de aprender mais sobre a relação verdadeira subjacente e de interpretar os efeitos, separados e conjuntos, produzidos pelas mudanças em determinadas variáveis importantes.

Mesmo quando não existe relação física entre as variáveis, podemos querer relacioná-las através de algum tipo de equação matemática. Enquanto a equação pode ser fisicamente sem sentido, pode, ainda assim, ser extremamente valiosa para predizer os valores de algumas variáveis a partir do conhecimento de outras variáveis.

Distinguiremos as variáveis em dois tipos principais: variáveis preditoras e variáveis respostas. Por variáveis preditoras entenderemos variáveis que podem ser ajustadas para um valor desejado ou então assumir valores que podem ser observados. Como resultado de alterações feitas nas variáveis preditoras, um efeito é transmitido a outras variáveis, as variáveis respostas.

Em geral estamos interessados em analisar como mudanças nas variáveis preditoras afetam os valores das variáveis respostas. Portanto, o objetivo é encontrar uma simples relação ou dependência da variável resposta com apenas uma ou com algumas poucas variáveis preditoras.

Estaremos interessados em relações da seguinte forma:

Variável resposta = função do modelo + erro aleatório

A função do modelo geralmente será conhecida, de forma especificada e envolverá as variáveis preditoras bem como os parâmetros a serem estimados a partir dos dados e, eventualmente, de outras fontes de informação.

O modelo linear de 1° grau, também conhecido como modelo de Regressão Linear Simples, seria da forma

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

onde:

$Y_i$ - valor observado para a variável dependente Y no i-ésimo nível da variável independente X;

$\beta_0$ - constante de regressão, que representa o intercepto da reta $\beta_0 + \beta_1 X_i$;

$\beta_1$ - coeficiente de regressão, que representa a variação de Y em função da variação de uma unidade da variável X;

$X_i$ - i-ésimo nível da variável independente X (i = 1,2,...,n);

$\varepsilon_i$ - erro associado à distância entre o valor observado $Y_i$ e o correspondente ponto na curva do modelo proposto, para o mesmo nível i de X.

Para que esse modelo possa ser aplicado, os erros devem satisfazer as seguintes suposições:

1. $E(\varepsilon_i) = 0$;
2. Normalidade;
3. Independência;
4. Homocedasticidade.

Para obtenção da equação estimada, utilizaremos o Método da Máxima Verossimilhança, que consiste em determinar os estimadores de máxima verossimilhança para os parâmetros $\beta_0$ e $\beta_1$, ou seja, os valores de $\widehat{\beta_0}$ e $\widehat{\beta_1}$ que maximizam a função de verossimilhança.

Consideremos uma amostra (Yi,Xi), i = 1, 2, ... , n e o modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Sob as suposições 2 e 4, temos que, condicionalmente a $\beta_0$, $\beta_1$e Xi, Yi são independentes e com distribuição

$$Yi \sim N(\beta_0 + \beta_1 X_i; \sigma^2)$$

Consequentemente, a função de densidade conjunta de Y1, Y2, ... , Yn pode ser escrita como

$$f(Y1, Y2, \dots , Yn \mid \beta_0 + \beta_1 X_i; \sigma^2)$$

Yi são independentes, logo

$$f(Y1, Y2, \dots , Yn \mid \beta_0 + \beta_1 X_i; \sigma^2) = \prod_{i=1}^{n} f(Y_i \mid \beta_0 + \beta_1 X_i; \sigma^2) \quad (1)$$

onde

$$f(Y_i|\beta_0 + \beta_1 X_i; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} * \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\} \quad (2)$$

que é a função densidade de uma variável com distribuição normal com média $\beta_0 + \beta_1 X_i$ e variância $\sigma^2$.

Substituindo (2) em (1), temos que

$$L(\beta_0, \beta_1, \sigma^2) = f(Y1, Y2, \ldots, Yn \mid \beta_0 + \beta_1 X_i; \sigma^2)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} * \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right\}$$

é a função de máxima verossimilhança do modelo.

Assim, sob o enfoque clássico, com o Método da Máxima Verossimilhança, os valores de $\beta_0$ e $\beta_1$ que maximizam esta função são

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Uma vez obtidas estas estimativas, podemos escrever a equação estimada como sendo

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$$

Sob o enfoque Bayesiano, obtemos a distribuição a posteriori dos parâmetros $\theta = (\beta_0, \beta_1, \sigma^2)$, tal que $P(\theta|\underline{Y})$ α $L(\theta|\underline{Y}) * P(\theta)$, onde $P(\theta)$ é a distribuição a priori de θ. Admitindo priori não-informativa (quando se espera que a informação dos dados seja dominante, no sentido de que a nossa informação a priori é vaga), os estimadores bayesianos para $\beta_0$ e $\beta_1$ coincidem com os estimadores de máxima verossimilhança.

Com o modelo ajustado, as suposições de 1 a 4 anteriores devem ser verificadas para a validação do mesmo. Chamamos de análise dos resíduos o conjunto de técnicas utilizadas para investigar a adequação do modelo de Regressão com base nos resíduos.

Quando a suposição de não-correlação entre os erros não é satisfeita, podemos recorrer a modelos mais complexos que acomodem essa propriedade dos erros. Por exemplo, no caso de séries temporais, podemos utilizar modelos dinâmicos para acomodar a dependência no tempo. No caso de dados espaciais, podemos usar métodos de Geoestatística para modelar dependência no espaço.

Para detalhes sobre ajuste e verificação das hipóteses de um modelo de regressão linear gaussiano ver Montgomery (2006).

# 3 Processos Estocásticos

## 3.1 Definição

Um processo estocástico é uma família de variáveis aleatórias definidas em um espaço de probabilidades (Ω, ∱ ,P), indexadas por elementos t ∈ **T** (espaço de índices ou parâmetros). Pode ser denotado como {Z(t): t ∈ **T**} , e as variáveis aleatórias como Z(t) ou $Z_t$ . Estamos definindo Ω como um conjunto arbitrário não vazio, ∱ como $\sigma$-álgebra de subconjunto de Ω e P como uma medida de probabilidade. (Marques,1996).

Se **T** é um conjunto enumerável, então o processo estocástico é dito um processo a tempo discreto, se **T** é um intervalo do conjunto dos números Reais, então é dito um processo a tempo contínuo. (Ross,1997).

Exemplos: Se Z(t) representa o número de usuários em uma fila de banco em um determinado instante, Z(t) é um processo a tempo contínuo, se representa o índice pluviométrico diário, é um processo a tempo discreto.

Z(t) pode também ser chamado de estado do processo em t. E o espaço de estados do processo são todos os valores possíveis de Z(t). (Ross, 1997).

No caso de um processo com espaço de estados discreto, Z(t) pode representar uma contagem, como, por exemplo, o número de chamadas telefônicas que chegam a uma central durante um período de duas horas. Já no caso de um processo com espaço de estados contínuo, Z(t) representa uma medida que varia continuamente, como temperatura, preço de um ativo financeiro, altura de ondas etc. (Morettin e Toloi, 2004).

Frequentemente, o índice t é interpretado como o tempo, neste caso podemos dizer que {Z(t): t ∈ **T**} é um processo que depende do tempo. Por exemplo, Z(t) pode ser o número total de clientes que entram em um supermercado no tempo t, ou o número de clientes no supermercado no tempo t, ou o número de vendas feitas no supermercado no tempo t. (Ross, 1997).

O índice t pode também ser interpretado como locais. Neste caso, podemos dizer que {Z(t): t ∈ **T**} é um processo que depende da localização, que é o nosso interesse para este projeto.

Neste estudo, iremos considerar um determinado tipo de processo estocástico: Processo Gaussiano.

## 3.2 Processo Gaussiano

Um processo {Z(t): t ∈ **T**} é dito gaussiano se as amostras $Z_1$ = Z($t_1$), $Z_2$ = Z($t_2$), ..., $Z_k$ = Z($t_k$) são variáveis aleatórias conjuntamente gaussianas para todo k e todas as escolhas de $t_1$, $t_2$, ..., $t_k$. (Marques, 1996).

Ou seja, um processo estocástico é dito gaussiano se Z($t_1$),..., Z($t_k$) tem uma distribuição Normal Multivariada para todo $t_1$, ..., $t_k$. (Ross, 1997).

Logo, a função de densidade de probabilidade conjunta das variáveis é:

$$f_{Z(t_1),Z(t_2),\dots,Z(t_k)}(z_1, z_2, \dots, z_k) = \frac{1}{2\pi^{\frac{k}{2}} \times \Delta^{\frac{1}{2}}} \times e^{-\frac{1}{2} \times (z-\mu)^T \times \Sigma^{-1} \times (z-\mu)}$$

Sejam:

$\mu = [\mu_1, \mu_2, \dots, \mu_k]$ é o vetor das médias

$\Sigma = \begin{bmatrix} C_Z(t_1,t_1) & \dots & C_Z(t_1,t_k) \\ \dots & \dots & \dots \\ C_Z(t_k,t_1) & \dots & C_Z(t_k,t_k) \end{bmatrix} = \{C_Z(t_i,t_j)\}_{i,j=1,\dots,k}$ é a matriz de covariâncias

e Δ é o determinante da matriz Σ.

Então, seguem algumas propriedades:

1. $f_{Z(t_1),Z(t_2),\dots Z(t_k)}(z_1, z_2, \dots, z_k) \geq 0$
2. $\int_{-\infty}^{\infty} f_{Z(t_1),Z(t_2),\dots Z(t_k)}(z_1, z_2, \dots, z_k)\, dz_1, z_2, \dots, z_k = 1$
3. $\int_{-\infty}^{\infty} z_i \times f_{Z(t_1),Z(t_2),\dots Z(t_k)}(z_1, z_2, \dots, z_k)\, dz_1, z_2, \dots, z_k = \mu_i\,, sendo\ 1 \leq i \leq k$
4. $\int_{-\infty}^{\infty} (z_i - \mu_i) \times (z_j - \mu_j) \times f_{Z(t_1),Z(t_2),\dots,Z(t_k)}(z_1, z_2, \dots, z_k)\, dz_1, z_2, \dots, z_k = C_Z(t_i, t_j),$ $sendo\ 1 \leq i \leq k\ e\ 1 \leq j \leq k.$
5. Dados um vetor $Z = \{z_1, z_2, \dots, z_k\}$ normalmente distribuído em k dimensões com vetor das médias $\mu$ de dimensão k e matriz de covariâncias, k × k, denominada Σ e $P$, uma matriz m × k. Então, P × Z = W, onde W=$\{w_1, w_2, \dots, w_m\}$ tem uma distribuição Gaussiana m-dimensional com vetor de médias de dimensão m igual a P × $\mu$ e matriz de covariâncias, m × m, igual a $\mathrm{P} \times \Sigma \times \mathrm{P}^{\mathrm{T}}$.
6. Dado $(z, z^*)^T \sim N\left((m, m^*)^T, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$, então

   $z \backslash z^* \sim N\left(m + \Sigma_{12}\Sigma_{22}^{-1}(z^* - z), \Sigma_{11} - \Sigma_{12}^T\Sigma_{22}^{-1}\Sigma_{21}\right)$. (Miller, 1962-64).

Podemos também definir a função de correlação como

$$r(t_i, t_j) = \frac{C_z(t_i,t_j)}{\sqrt{C_z(t_i,t_i)\, C_z(t_j,t_j)}}.$$

Alguns exemplos de processos gaussianos são o Movimento Browniano Padrão, Processo de Wilner, Processo de Ornstein-Uhlenbeck. (Ross, 1997).

# 4 Estatística Espacial

Entende-se como Estatística Espacial o ramo da estatística que, na análise dos dados, leva em consideração a localização.

Diversas áreas têm percebido a importância da aplicação da Estatística Espacial, entre elas: climatologia, ecologia, saúde ambiental, marketing imobiliário, demografia etc.

Os dados espaciais podem ser classificados em:

- Dados geoestatísticos: Os dados são medidas obtidas em pontos do espaço. O objetivo nestes modelos é analisar a relação das medidas com a proximidade dos pontos e prever uma medida em um ponto a partir de medidas já observadas de pontos próximos;
- Dados de áreas: Dados em áreas geográficas limitadas. São classificados desta forma quando temos um valor por área, não tendo a localização exata dos eventos;
- Dados de pontos: Nos modelos de dados de pontos, são dadas localizações aleatórias de um determinado evento, ou seja, locais onde houve a ocorrência do evento. A análise destes modelos tem como objetivo determinar se há dependência da localização com a ocorrência do evento.

Neste projeto, consideraremos os dados geoestatísticos.

## 4.1 Modelos de dados geoestatísticos

É importante definirmos {Y(s): s ∈ D} como um processo estocástico, em que D é um subconjunto fixado de um espaço Euclidiano r-dimensional. Para situações em que r>1, temos um processo espacial.

Podemos supor que a covariância entre duas variáveis aleatórias depende da distância entre as localizações delas. Uma associação freqüentemente usada é o modelo exponencial. Neste caso, a covariância entre medidas de duas localizações vai depender da distância entre elas: $COV\left[Y(s_i), Y\left(s_{i'}\right)\right] = C\left(d_{ii'}\right) = \sigma^2 e^{-\phi d_{ii'}}$, para $i \neq i'$, sendo $d_{ii'}$ a distância entre $s_i$ e $s_{i'}$ e $\sigma^2$ e $\phi$ parâmetros positivos chamados, respectivamente, de patamar parcial e parâmetro de decaimento e 1/ $\phi$ é chamado de parâmetro de amplitude. Para $i = i'$, $C\left(d_{ii'}\right) = VAR[\, Y(s_i)] = \sigma^2 + \tau^2$, onde $\tau^2 > 0$ é chamado de efeito pepita.

Como distribuição conjunta, é conveniente assumir uma distribuição Gaussiana para os dados. Suponha $\{Y(s_i)\}$, i=1,...,n, então: $Y|\mu, \theta \sim N_n(\mu, \Sigma(\theta))$, onde $\mu$ é a média, $\theta = (\tau^2, \sigma^2, \phi)^T$, $COV\left[Y(s_i), Y\left(s_{i'}\right)\right] = (\Sigma(\theta))_{ii'} = \sigma^2 e^{-\phi d_{ii'}} + \tau^2 I_{(i=i')}$.

É usual, na análise dos modelos de dados geoestatísticos, assumirmos estacionariedade para o processo {Y(s): s ∈ D}. Podemos assumir três tipos de estacionariedade:

- Estritamente estacionário: Se dado n≥1 e dado h ∈ $\Re^r$, temos o conjunto de n localizações: $\{s_1, \ldots, s_n\}$, e a distribuição de $\{Y(s_1), \ldots, Y(s_n)\}$ é a mesma que a distribuição de $\{Y(s_1 + h), \ldots, Y(s_n + h)\}$;
- Fracamente estacionário: Se $\mu(s) \equiv \mu$, sendo $\mu(s)$ a média do processo, e se COV[ Y(s), Y(s+h) ]= C(h), para todo h ∈ $\Re^r$. Como assumiremos média zero, a condição é que a função de covariância dependa apenas do vetor de separação h;
- Intrinsecamente estacionário: Assumimos E[Y(s+h)–Y(s)]=0 e definimos E[Y(s+h) –Y(s)]² = VAR[Y(s+h)–Y(s)] = $2\gamma(h)$, onde a função $2\gamma(h)$ é chamada de variograma e $\gamma(h)$ é chamada de semi-variograma.

Desenvolvendo $2\gamma(h)$, encontramos o variograma em função das covariâncias: $2\gamma(h) = 2$ [ C(0) - C(h) ], C(h) pode ser chamado de covariograma. De forma intuitiva, podemos dizer que quando $\|h\| \to \infty$, então $C(h) \to 0$. Como $2\gamma(h)$ e $C(h)$ são inversamente proporcionais, temos que quanto maior h, menor o covariograma e maior o variograma.

Se o variograma depende de h apenas por meio de seu comprimento $\|h\|$, então dizemos que o processo é isotrópico, senão é dito anisotrópico. Se um processo é isotrópico e intrinsicamente estacionário, então é dito homogêneo.

São alguns candidatos a semi-variogramas para processos isotrópicos, com t=$\|h\|$:

- Linear: $\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t, \text{se } t > 0,\ \tau^2 > 0, \sigma^2 > 0 \\ 0, \text{caso contrário} \end{cases}$
- Esférico: $\gamma(t) = \begin{cases} \tau^2 + \sigma^2, \text{se } t \geq 1/\phi \\ \tau^2 + \sigma^2\left\{\frac{3\phi t}{2} - \frac{1}{2}(\phi t)^3\right\}, \text{se } 0 < t \leq 1/\phi \\ 0, \text{caso contrário} \end{cases}$
- Exponencial: $\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t)), \text{se } t > 0 \\ 0, \text{caso contrário} \end{cases}$
- Gaussiano: $\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 t^2)), \text{se } t > 0 \\ 0, \text{caso contrário} \end{cases}$
- Matérn: $\gamma(t) = \begin{cases} \tau^2 + \sigma^2\left(1 - \frac{(2\sqrt{\nu}t\phi)^\nu}{2^{\nu-1}\Gamma(\nu)}\mathrm{K}_\nu\left(2\sqrt{\nu}t\phi\right)\right), \text{se } t > 0 \\ 0, \text{caso contrário} \end{cases}$

onde $\nu$ é um parâmetro de controle da suavização, $\phi$ é um parâmetro espacial e $\mathrm{K}_\nu$ é a função Bessel modificada de ordem $\nu$.

O modelo de semi-variograma é escolhido achando o semi-variograma na forma empírica e comparando com os modelos existentes, analisando o que melhor se adequa. O semi-variograma na forma empírica pode ser encontrado pelo Estimador Clássico ou pelo Estimador Módulo:

Estimador Clássico (considera a soma do quadrado das diferenças dos valores) : $\hat{\gamma}(t) = \frac{1}{2N(t)}\sum_{(s_i,s_j)\in N(t)}[Y(s_i) - Y(s_j)]^2$

Estimador Módulo (considera a soma dos módulos das diferenças dos valores): $\hat{\gamma}(t) = \frac{1}{2N(t)} \sum_{(s_i,s_j)\in N(t)} \left|Y(s_i) - Y\left(s_j\right)\right|$

onde N(t) é o conjunto de pares de pontos tal que $t = \left\|s_i - s_j\right\|$.

A análise do modelo que melhor se adequa pode ser feita visualmente e pelo método de mínimos quadrados.

## 4.2 Dados Espaciais Univariados

O modelo univariado é escrito como: Y(s) = $\mu$(s) + w(s) + $\epsilon$(s), onde $\mu$(s)=$x^T(s)\beta$ é a média e o resíduo é separado em duas partes: espacial w(s), sendo (w(s)| $\sigma^2, \phi$) ~ processo gaussiano PG(0; $\sigma^2$H($\phi$)) e não espacial $\epsilon$(s), sendo $\epsilon$~N(0, $\tau^2$).

O problema agora se resume a estimar os parâmetros. No caso do efeito pepita, por exemplo, temos: $\Sigma = \sigma^2 H(\phi) + \tau^2 I$, onde H é a matriz de correlação: $H_{ij} = \rho(s_i - s_j, \phi)$. Assumindo $\theta = (\beta, \sigma^2, \tau^2, \phi)^T$, em uma análise Bayesiana, temos como posteriori:

$$p(\theta|y) \propto f(y|\theta)p(\theta)$$

onde:

$Y|\theta \sim N(X\beta, \sigma^2 H(\phi) + \tau^2 I)$

$p(\theta) = p(\beta)p(\sigma^2)p(\tau^2)p(\phi)$.

Para fazermos afirmações inferenciais sobre os parâmetros separadamente, temos que obter distribuições marginais a posteriori. Por exemplo, uma estimativa pontual ou um intervalo de credibilidade para $\beta$ surge a partir de:

$$p(\beta|y) = \iiint p(\beta, \sigma^2, \tau^2, \phi|y)\, d\sigma^2 d\tau^2 d\phi$$

$$\propto p(\beta) \iiint f(y|\theta)p(\sigma^2)p(\tau^2)p(\phi)\, d\sigma^2 d\tau^2 d\phi$$

Na prática, não há forma fechada para a integração acima, logo teremos que recorrer a métodos computacionais para obtenção da posteriori. Em particular, utilizaremos os métodos MCMC (Monte Carlo via Cadeia de Markov), vide seção 5.5.

A expressão $Y|\theta$ pode ser reformulada como um modelo hierárquico, escrevendo Y condicional não só de $\theta$, mas também de W=$(\mathrm{w}(s_1), \ldots, \mathrm{w}(s_n))^T$:

$$Y|\theta, W \sim N(X\beta + W, \tau^2 I)$$

onde: $W|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi))$

Então, a posteriori pode ser escrita como:

$$p(\theta, W|y) \propto f(y|\theta, W)p(W|\theta)p(\theta)$$

Se quisermos prever um $y(s_0)$ desconhecido, dado $x(s_0)$, então é necessário encontrar a distribuição preditiva:

$$p\big(y(s_0)|y, X, x(s_0)\big) = \int p\big(y(s_0), \theta|\mathrm{y}, \mathrm{X}, x(s_0)\big)\, \mathrm{d}\theta = \int p(y(s_0)|\mathrm{y}, \theta, x(s_0)p(\theta|y, X)d\theta$$

Na prática, também não há solução analítica e uma possibilidade é o uso de métodos MCMC.

## 4.3 Dados Espaciais Multivariados

Denotemos W(s) como o vetor de variáveis aleatórias (px1) medidas na localização s.

Vamos denotar Y(s) como a variável resposta e x(s) como o vetor das covariáveis medidas na localização s. Separemos as localizações em 4 partes, ou seja, s pertence a um dos grupos abaixo:

- $S_Y$ é o conjunto das localizações onde apenas Y(s) foi observado,
- $S_X$ é o conjunto das localizações onde apenas as covariáveis foram observadas,

- $S_{YX}$ é o conjunto das localizações onde a variável resposta e as covariáveis foram observadas e
- $S_U$ é o conjunto das localizações onde nenhuma medida foi observada.

Assim, podemos pensar em três objetivos:

- Encontrar W(s) quando s ∈ $S_X$, usando a interpolação,
- Encontrar W(s) quando s ∈ $S_U$, usando a predição e
- Encontrar W(s) através da relação com X(s) e outras covariáveis informativas de s, denominadas U(s), usando a Regressão Espacial.

Como vimos no capítulo 3, na Regressão normalmente utilizada, os pares (X($S_i$), Y($S_i$)), i=1,...,n são independentes, suprimindo U($S_i$). No contexto espacial, isto não acontece, portanto, nosso objetivo neste trabalho é entender a utilização da Regressão Espacial, pois devemos considerar U(s), já que a dependência espacial existe. Portanto, queremos determinar o valor esperado de Y(s) dados X(s) e U(s): E[Y(s)|X(s),U(s)].

Suponha: $W(s) = \binom{X(s)}{Y(s)} \sim N(\mu(s), T)$.

Denote X como o vetor considerando as localizações s em que as covariáveis foram observadas, X' o vetor considerando as localizações s em que as covariáveis não foram observadas, Y como o vetor considerando as localizações s em que as variáveis respostas foram observadas e Y' o vetor considerando as localizações s em que as variáveis respostas não foram observadas.

Na implementação do amostrador de Gibbs atualiza-se X', Y' e os parâmetros utilizados dados X e Y, de forma a montar o modelo a ser usado.

Para o problema de Regressão, a distribuição de interesse é $p(E[Y(s_0)|x(s_0)]|x(s_0), y, x)$, onde $x(s_0)$ é observado e $y(s_0)$ não. Suponha $\mu(s) = (\mu_{1,}\mu_2)^T$, $\mu_2(s) = \alpha^T U(s)$. Então, para o par $\big(X(s), Y(s)\big), p(y(s)|x(s), \beta_0, \beta_1, \sigma^2)$ é $N(\beta_0 + \beta_1 x(s), \sigma^2)$. Portanto, $E[Y(s)|x(s)] = \beta_0 + \beta_1 x(s)$, onde:

$$\beta_0 = \mu_2 - \frac{T_{12}}{T_{11}}\mu_{1,} \quad \beta_1 = \frac{T_{12}}{T_{11}} \quad e \quad \sigma^2 = T_{22} - \frac{{T_{12}}^2}{T_{11}}$$

Será necessário, agora, rearranjar as componentes de W de forma a simplificar o cálculo:

$$\binom{X}{Y} \sim \left(N\binom{\mu_1 1}{\mu_2 1}, T \otimes H(\phi)\right)$$

Podemos assumir T com distribuição a priori Wishart, e para que seja possível a utilização da estatística Bayesiana, precisaremos assumir prioris também para $\mu_1, \mu_2 \, e \, \phi$. Para $(\mu_1, \mu_2)$ assumiremos como priori a normal bivariada. A priori que assumiremos para $\phi$, depende da escolha de $\rho(h, \phi)$. E, então, será usado amostrador de Gibbs para simular as distribuições posterioris necessárias. Porém a condicional de $\phi$ depende das entradas de H, que não tem sua forma fechada, e, portanto, pode ser utilizado Metropolis para esta atualização.

Portanto, dado $s_0$ uma nova localização onde gostaríamos de prever variáveis de interesse, o primeiro passo necessário será modificar a matriz H($\phi$) pela nova matriz H*:

$$H^*(\phi) = \begin{pmatrix} H(\phi) & h(\phi) \\ h(\phi)^T & \rho(0, \phi) \end{pmatrix}$$

onde $h(\phi)$ é o vetor composto por $\rho(s_0 - s_j, \phi)$, j=1,2,...,n.

Então, segue que:

$$W^* \equiv \big(W(s_0), \dots, W(s_n)\big) \sim N\left(1_{n+1} \otimes \binom{\mu_1}{\mu_2}, H^*(\phi) \otimes T\right)$$

Agora, a distribuição preditiva pode ser obtida pelas marginais dos parâmetros:

$$p(y(s_0)|y, x) = \int p(y(s_0)|x(s_0), y, x, \mu, T, \phi) p\,(\mu, T, \phi|x(s_0), y, x)$$

## 4.4 Dados Espaciais Bivariados – Distribuição Condicional

Para modelar dados bivariados utilizando distribuições condicionais, considere para o modelo Y(s) = $\mu$(s) + v(s) + $\epsilon$(s):

$$v(s) = A \times w(s)$$

onde $w_j(s)$ é um Processo Gaussiano com média zero, variância 1 e matriz de correlação $\rho_j$ e A é uma matriz triangular inferior.

Sobre v(s), podemos escrever, sem perda de generalidade:

$$p\big(v(s)\big) = p(v_1(s))\, p(v_2(s)|v_1(s))$$

sendo:

- p$(v_1(s))$ ~ N(0,$T_{11}$), portanto, $v_1(s) = \sqrt{T_{11}}w_1(s) = a_{11}w_1(s)$, $a_{11} > 0$
- $p\big(v_2(s)|v_1(s)\big) \sim N\left(\frac{T_{12}v_1(s)}{T_{11}}; T_{22} - \frac{T_{12}^2}{T_{11}}\right) = N\left(\frac{a_{21}}{a_{11}}v_1(s); a_{22}^2\right)$.

  com:

- $v_2(s_i) = \frac{a_{21}}{a_{11}}v_1(s_i) + a_{22}w_2(s_i)$
- $v_1(s_i) = \sigma_1 w_1(s)$

  Logo: $v_2(s)|v_1(s) = \alpha v_1(s) + \sigma_2(s)w_2(s)$

Podemos calcular $(a_{11}, a_{12}, a_{22})$ e $(T_{11}, T_{12}, T_{22})$ através da parametrização $(\alpha, \sigma_1, \sigma_2)$:

- $a_{11} = \sigma_1,\ a_{21} = \alpha\sigma_{1,}\ a_{22} = \sigma_2$
- $T_{11} = \sigma_1^2,\ T_{12} = \alpha\sigma_1^2,\ T_{22} = \alpha^2\sigma_1^2 + \sigma_2^2$

Suponha $\phi_j$ um parâmetro associado à função de correlação $\rho_j$, então a distribuição de v depende de T e $\phi$. Vamos assumir uma priori conjunta p(T, $\phi$) = p(T)p( $\phi$), ou seja, p($\sigma$,$\alpha$, $\phi$) = p($\sigma$,$\alpha$)p( $\phi$). A priori padrão associada a T é uma

Wishart Invesa, associada a $\sigma^2$ é uma Gama Inversa e associada a $\alpha$ é uma Normal.

Se:

- $Y_1(s) = X_1^T\beta_1 + v_1(s) + \epsilon_1(s)$
- $Y_2(s) = X_2^T\beta_1 + v_2(s) + \epsilon_2(s)$

então o modelo condicional será:

- $Y_1(s) = X_1^T\beta_1 + \sigma_1 w_1(s) + \tau_1 u_1(s)$
- $Y_2(s)|Y_1(s) = X_2^T\beta_1 + \alpha Y_1(s) + \sigma_2 w_2(s) + \tau_2 u_2(s)$

onde $u_1(s), u_2(s) \sim N(0,1)$.

A distribuição a posteriori dos parâmetros não tem forma fechada e, portanto, é necessário o uso de alguma aproximação computacional para a posteriori. Em particular, faremos uso dos métodos MCMC.

Para detalhes sobre Estatística Espacial ver Gelfand (2003).

## 4.5 Monte Carlo via Cadeia de Markov

Como não é possível fazer sumarização da distribuição a posteriori de interesse neste trabalho de forma analítica, faremos uso de métodos de simulação estocástica para obter amostras da densidade.

Na última década a inferência Bayesiana vem experimentando um grande avanço devido à introdução de métodos de Monte Carlo via Cadeias de Markov (MCMC) e também devido à disponibilidade de computadores velozes. O MCMC é uma técnica poderosa que vem permitindo a análise de modelos altamente estruturados. Nesse contexto os métodos de simulação estocástica mais utilizados são o amostrador de Gibbs e o Metropolis-Hastings.

Paralelamente ao desenvolvimento de técnicas de simulação estocástica, surgiram também programas de computadores que geram amostras da posteriori de modelos altamente estruturados. O projeto WinBUGS1 (Spiegelhalter et al., 2002) vem permitindo a implementação de modelos hierárquicos relativamente complexos e, particularmente, de modelos espaciais. Outro exemplo de programa para obtenção da posteriori de modelos espaciais é o geo-R2 (Ribeiro Jr and Diggle, 2001).

O método de integração por Monte Carlo é utilizado para aproximar integrais que são difíceis ou impossíveis de serem calculadas analiticamente, particularmente quando a dimensão do problema é grande.

Esta seção introduz os métodos de simulação estocástica via cadeias de Markov (MCMC), os quais serão usados para obter amostras de densidades a posteriori complexas. Para uma introdução detalhada aos métodos MCMC ver Gilks, Richardson e Spiegelhalter (1996), Gamerman (1997) e Robert e Casella (2004).

### 4.5.1 Cadeias de Markov

Uma cadeia de Markov é uma coleção de variáveis aleatórias (vetores aleatórios) $\{X \in \Omega | i \in M\}$, onde usualmente $M = N$. A evolução da cadeia de Markov no espaço $\Omega \subset R^p$ é dada pelo núcleo de transição.

$$P(x, A) = P(X_{i+1} \in A | X_i = x, X_j, j < i) = P(X_{i+1} \in A | X_i = x) \ (x \in \Omega, A \subset \Omega) \ (3)$$

Isto quer dizer que uma cadeia de Markov é um processo estocástico, em que, dado o estado presente, passado e futuro são independentes.

Em geral, o núcleo de transição tem um componente contínuo e um componente discreto para alguma função $\mathrm{p}: \Omega \times \Omega \rightarrow (0, \infty)$, sendo expressado por

$$P(x, dy) = P(x, y)dy + r(x)I_{dy}(x) \ (4)$$

onde $P(x, x) = 0$ e $r(x) = 1 - \int_{\Omega} P(x, y)dy$. Do mesmo modo, a transição de $x$ para $y$ ocorre de acordo com $P(x, y)$, e a transição de $x$ para $x$ ocorre com probabilidade

$r(x)$. Da equação (3) segue que o núcleo de transição proporciona a distribuição de $X_{i+1}$ dado que $X_i = x$. O núcleo de transição n-passos a frente é dado por:

$$P^{(n)}(x, A) = \int_{\Omega} P^{(n-1)}(y, A) P(x, dy)$$

onde $P^{(1)}(x, dy) = P(x, dy)$ e $P(x, A) = \int_A P(x, dy)$.

Sob certas condições de regularidade, que serão mencionadas a seguir, a distribuição dada pela n-ésima iteração do núcleo de transição converge a uma distribuição invariante quando $n \to \infty$. A condição de invariância estabelece que se $X_i$ tem distribuição invariante, então todos os subsequentes elementos da cadeia também têm distribuição invariante.

### 4.5.2 O algoritmo de Metropolis-Hastings

O algoritmo de Metropolis foi apresentado inicialmente por Metropolis, Rosenbluth, Rosenbluth, Teller e Teller (1953) e generalizado por Hastings (1970), resultando no algoritmo de Metropolis-Hastings. Esse método é usado geralmente quando é difícil gerar amostras da distribuição a posteriori, $\pi(\theta)$. Neste caso, são gerados valores do parâmetro a partir de uma distribuição proposta $q(\theta|\theta^{(i-1)})$ que são aceitos ou não com certa probabilidade.

Para descrever o algoritmo, suponha que a distribuição de interesse é $\pi(\theta)$ e que a distribuição proposta é $q(\theta|\theta^{(i-1)})$, a qual será usada para obter $\theta^*$ dado o valor atual $\theta^{(i-1)}$ e seja $\alpha_{MH}(\theta^{(i-1)}, \theta^*)$ a probabilidade de aceitação.

$$\alpha_{MH}\left(\theta^{(i-1)}, \theta^*\right) = \min\left\{\frac{\pi(\theta^*) q(\theta^{)i-1)}|\theta^*)}{\pi(\theta^{(i-1)}) q(\theta^*|\theta^{(i-1)})}, 1\right\}$$

De maneira algorítmica, os valores simulados podem ser obtidos a partir do seguinte procedimento recursivo.

#### 4.5.2.1 Algoritmo M-H

1) Especificar um valor inicial $\theta^{(0)}$ tal que $\pi(\theta^{(0)}) > 0$ e fazer $i = 0$.

2) Gerar uma proposta $\theta^{*}\sim q(\theta|\theta^{(i-1)})$.

3) Gerar $u\sim\mathcal{U}(0,1)$.

4) Fazer

$$\theta^{(i)} = \begin{cases} \theta^{*}, se\ u \leq \alpha_{MH}(\theta^{(i-1)}, \theta^{*}) \\ \theta^{(i-1)}, em\ outro\ caso \end{cases}$$

5) Fazer $i = i + 1$, voltar a 2 e continuar o procedimento até alcançar a convergência.

### 4.5.3 O amostrador de Gibbs

O amostrador de Gibbs é um caso especial do algoritmo M-H que permite gerar uma amostra da distribuição a posterior $\pi(\theta)$i desde que as condicionais completas estejam disponíveis para amostragem. Uma introdução ao amostrador de Gibbs é dada por exemplo em Gamerman (1997) e Gelman et al. (2003). Já que o amostrador de Gibbs é simples e amplamente usado, não é necessariamente o procedimento mais eficiente na solução de um problema. Assim, em casos onde o amostrador de Gibbs não é a única aproximação possível, a simplicidade da implementação é uma vantagem que compensa sua ineficiência.

Para descrever o algoritmo suponha que a distribuição de interesse é a distribuição $\pi(\theta)$ onde $\theta = (\theta_1, \dots, \theta_d)'$. Cada $\theta_i$ pode ser um escalar ou um vetor. Considere também que todas as condicionais completas $\pi_i(\theta_i|\theta_{-i})$ estejam disponíveis e que se sabe gerar amostras de cada uma delas. Portanto, o esquema de amostragem é dado por:

#### 4.5.3.1 Algoritmo de Gibbs

1) Especificar um valor inicial $\theta^{(0)}$ e fazer $i = 0$.
2) Dado $\theta^{(i-1)}$, o próximo valor é obtido por simulação

$$\theta_1^{(i)} \sim \pi_1(\theta_1 | \theta_2^{(i-1)}, \dots, \theta_d^{(i-1)})$$

$$\theta_2^{(i)} \sim \pi_2(\theta_2 | \theta_1^{(i)}, \theta_3^{(i-1)} \dots, \theta_d^{(i-1)})$$

$$\vdots$$

$$\theta_d^{(i)} \sim \pi_d(\theta_d | \theta_1^{(i)}, \dots, \theta_{d-1}^{(i)})$$

Note que aqui o processo de atualização segue uma ordem fixa. Isto não é necessário. A ordem pode ser aleatoriamente permutada a cada passo.

3) Fazer $i = i + 1$, voltar a 2 e continuar o procedimento até alcançar a convergência.

# 5 Aplicação

Foram disponibilizados, pelo site do INPE – Instituto Nacional de Pesquisas Espaciais, os valores da temperatura, em graus Celsius, e da umidade relativa do ar de 38 estações no Ceará, medidas em 26/09/2008 às 18 horas. Serão apresentadas neste capítulo aplicações destes dados.

## 5.1 Análise de Regressão

Apresentaremos uma análise de Regressão Simples como uma análise preliminar para esses dados.

Utilizaremos um modelo de regressão cuja variável resposta Y é a temperatura e a variável explicativa X é a umidade.

### 5.1.1 Testes de Normalidade dos Resíduos

A normalidade dos resíduos é uma suposição essencial para que os resultados do ajuste do modelo sejam confiáveis. Podemos verificar essa suposição por meio de testes tais como Shapiro-Wilk, Anderson-Darling e Kolmogorov-Smirnov. Utilizando o teste de Shapiro-Wilk, obtemos o p-valor 0,0887, o que nos leva a uma aceitação, ao nível de significância $\alpha_0$< 0,0887, da hipótese de que os resíduos têm distribuição normal.

Para uma análise gráfica, espera-se que os pontos do gráfico dos quantis amostrais dos resíduos versus os quantis teóricos dos resíduos sigam o comportamento de uma reta. Observamos na Figura 1 uma situação satisfatória. Portanto, existem indícios de que os erros são normalmente distribuídos.

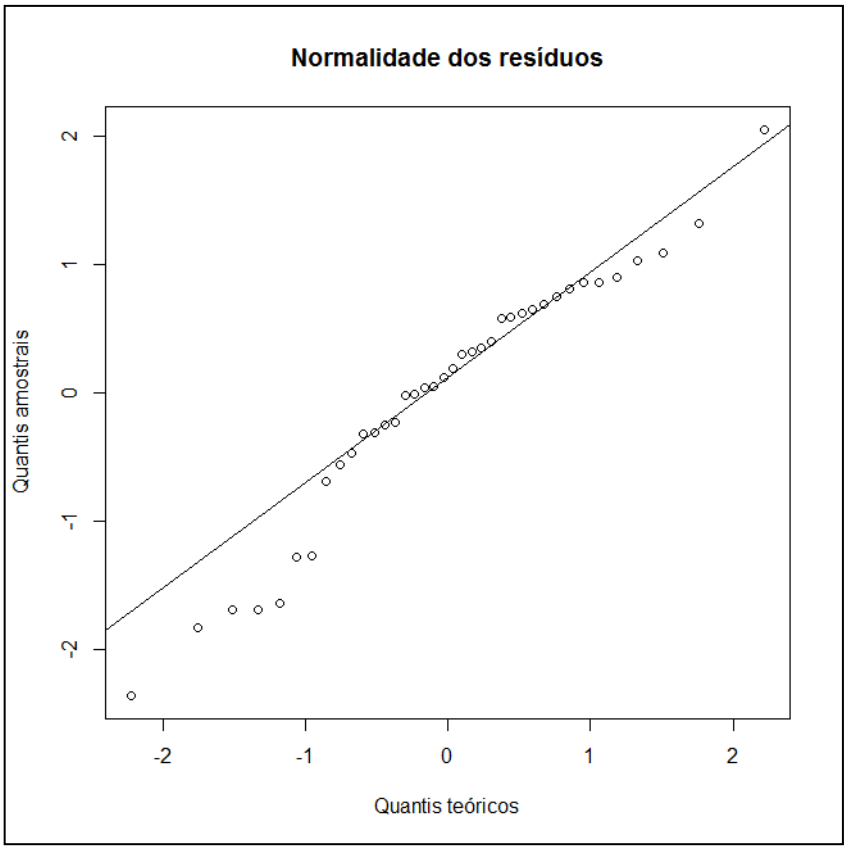



**Figura 1: Gráfico dos quantis amostrais dos resíduos versus os quantis teóricos dos resíduos para verificação da normalidade dos resíduos**

### 5.1.2 Média dos resíduos igual a zero (E($\varepsilon_i$) = 0)

Utilizando o Teste t de Student, obtemos o p-valor 1, o que nos leva a aceitar a hipótese de que a média dos resíduos é 0.

Observando a Figura 2, percebemos também que os resíduos estão com a média centrada no 0.

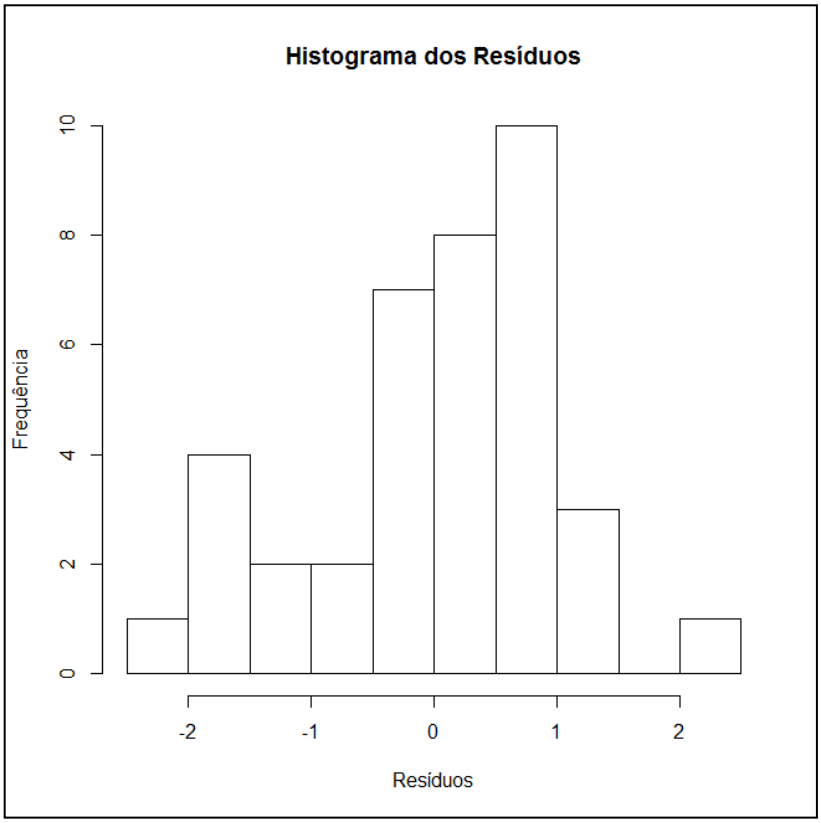



**Figura 2: Histograma dos Resíduos**

### 5.1.3 Teste de independência dos resíduos

A presença de autocorrelação (dependência) nos resíduos de um modelo de Regressão pode ser verificada através do teste de Durbin-Watson. Adotando este teste, obtemos o p-valor 0,04788, o que nos leva, ao nível de significância $\alpha_0$=0,05, à rejeição da hipótese de que os resíduos são não-correlacionados. Portanto, os resíduos apresentam dependência.

### 5.1.4 Teste de homocedasticidade dos resíduos

Homocedasticidade é o termo usado para designar variância $\sigma^2$ constante dos erros $\varepsilon_i$ para diferentes observações. Caso a suposição de homocedasticidade não seja válida, os erros padrões dos estimadores são incorretos, invalidando a inferência estatística, e não podemos dizer que os estimadores têm variância mínima para $\beta_0$ e $\beta_1$.

Podemos verificar a homocedasticidade dos resíduos por meio de testes tais como Breusch-Pagan e Goldfeld-Quandt. Adotando o teste de Breusch-Pagan, obtemos o p-valor 0,001051, o que nos leva, a qualquer nível de significância $\alpha_0$> 0,001051, à rejeição da hipótese de que variância dos resíduos é constante.

Outra técnica utilizada para verificar a homocedasticidade é o gráfico dos resíduos versus valores ajustados. Para diagnosticá-la, tentamos encontrar no gráfico alguma tendência. Como os pontos não estão aleatoriamente distribuídos em torno do 0, apresentando um comportamento (tendência), temos indícios de que a variância dos resíduos não é constante, ou seja, é heterocedástica.

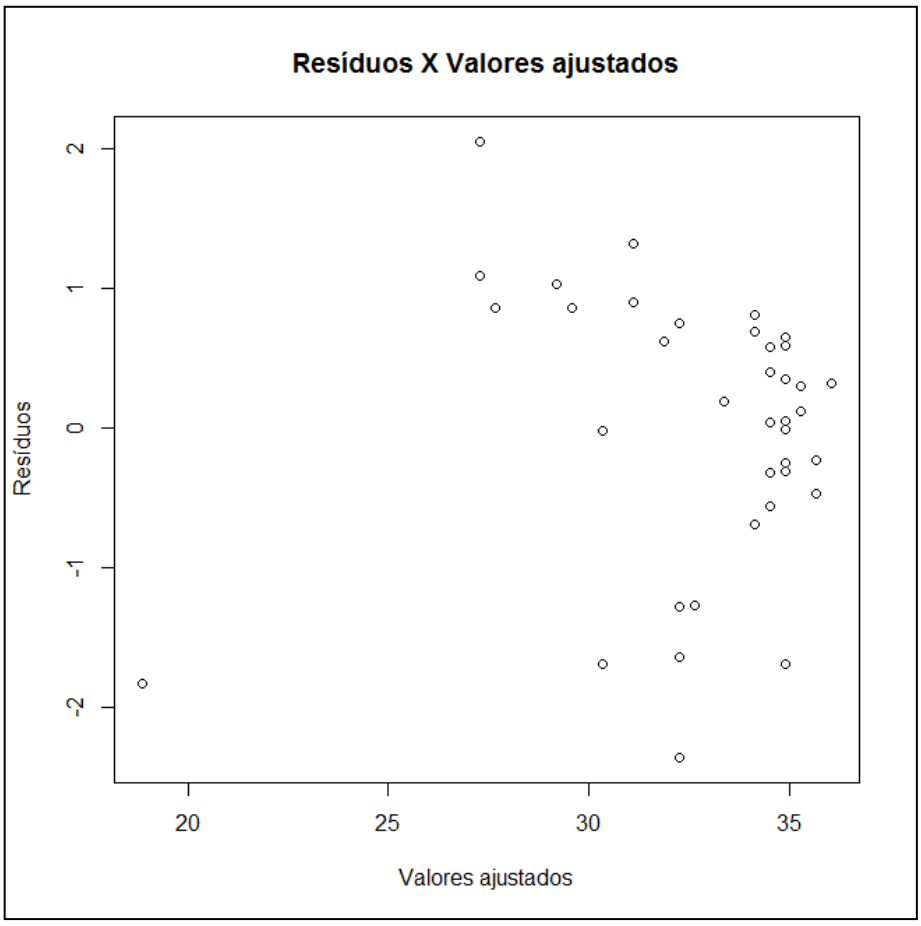



**Figura 3: Gráfico dos resíduos versus os valores ajustados para verificação da homocedasticidade**

Verificamos, então, que as suposições 3 e 4 da análise de resíduos não são satisfeitas: os resíduos não são independentes e a variância dos resíduos não é constante. Portanto, a Regressão Linear Simples não é um método adequado para modelar a relação existente entre temperatura e umidade relativa do ar.

## 5.2 Análise Exploratória de Dados

Nesta seção os dados serão analisados de forma a caracterizar as relações existentes entre as informações disponíveis.

Espera-se que a temperatura seja maior quanto menor for a umidade e vice-versa, pois a umidade é a razão da pressão parcial de vapor de água do ar e a pressão de vapor de saturação, e como a temperatura é diretamente proporcional à pressão de vapor de saturação, é, então, inversamente proporcional a umidade.

Vamos analisar agora se esta tendência pode ser percebida nestes dados.

Inicialmente, através da Figura 4, analisaremos um gráfico da temperatura em função da umidade.

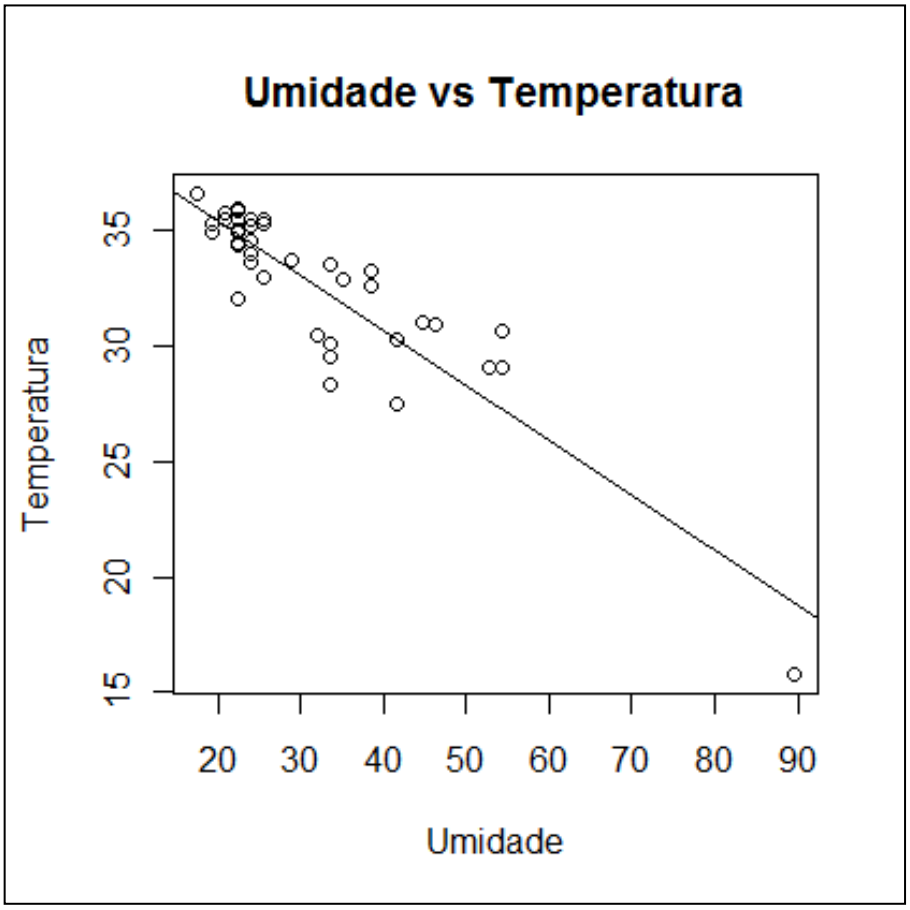



**Figura 4: Umidade versus Temperatura**

Na Figura 4, é perceptível um ponto cujo valor da umidade e da temperatura fogem do comum. Este ponto refere-se à estação de Pacajus. Vamos retirá-lo do nosso conjunto de dados e, através da Figura 5, verificar se é necessário retirá-lo para análises futuras.

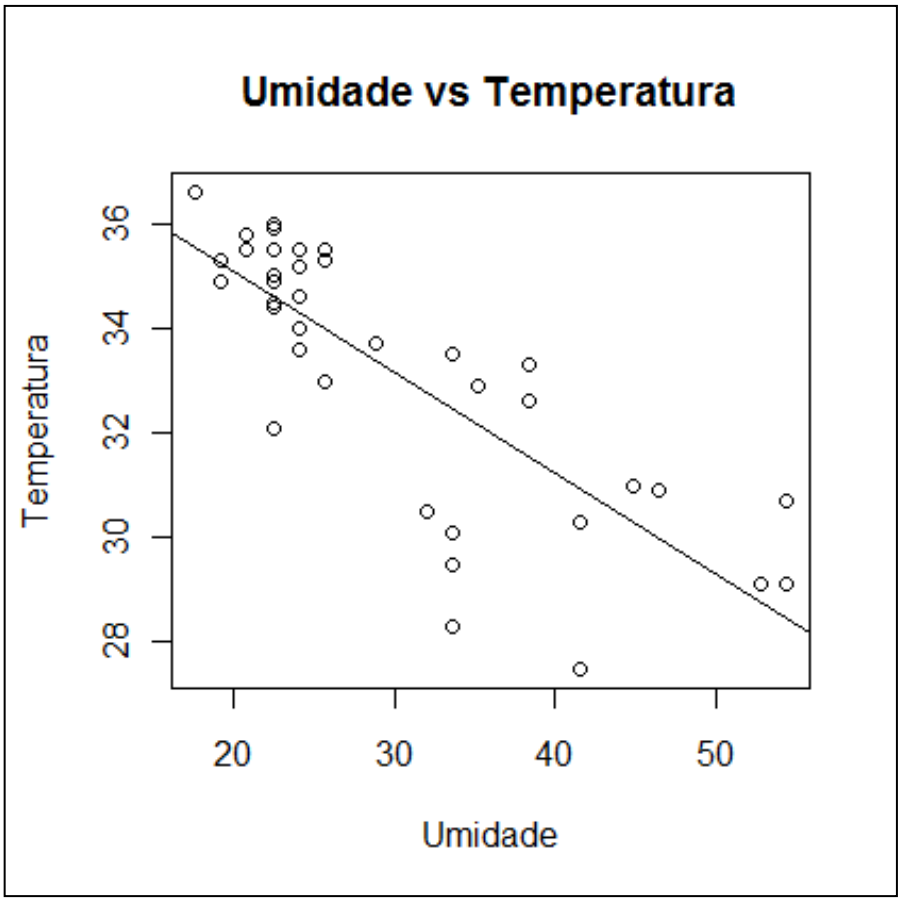



**Figura 5: Umidade versus temperatura sem as medidas na estação de Pacajus**

A inclinação da reta não muda e é, portanto, melhor manter a estação de Pacajus, pois apesar de afastado dos demais pontos, o ponto não é considerado um ponto de inclinação.

Podemos perceber que a reta ajustada decai no eixo y (temperatura) conforme cresce no eixo x (umidade). Portanto, a temperatura está diminuindo quando a umidade está aumentando, como o esperado.

Os valores da temperatura deste conjunto de dados variam de 15,80º a 36,6º, têm média amostral 32,68º e mediana 33,65º. Os valores da umidade variam de 17,60 a 89,60, têm média amostral 31,75 e mediana 25,60.

O próximo passo será analisar, através da Figura 6, os histogramas da temperatura e da umidade.

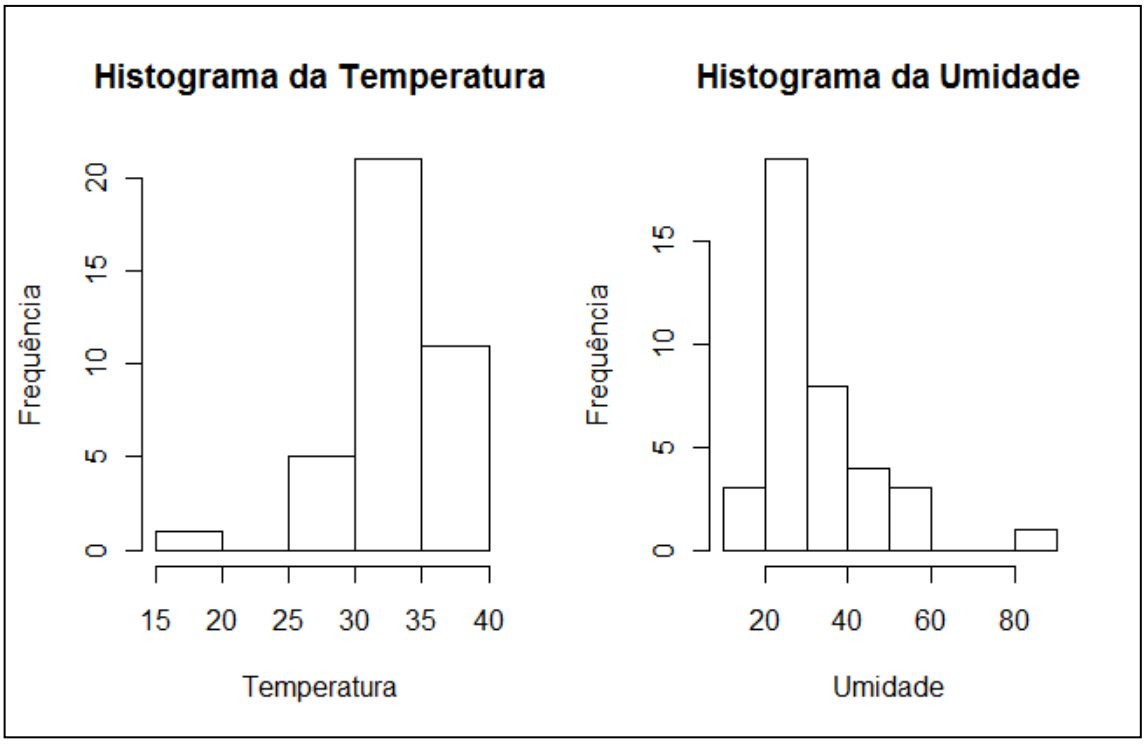



**Figura 6: Histograma da Temperatura e Histograma da Umidade**

Pode-se verificar que o Histograma da Temperatura apresenta assimetria para a direita, enquanto que o Histograma da Umidade apresenta assimetria para a esquerda, ou seja, os valores da temperatura e da umidade comportam-se de forma inversa. Podemos perceber também que nos dois histogramas há pouca dispersão, os valores concentram-se em pequenos intervalos.

Além da temperatura variar de forma inversa à umidade, também, intuitivamente, varia de acordo com a sua localização, pois locais mais próximos tendem a ter temperaturas mais parecidas e vice-versa. Ou seja, quando as distâncias são menores, diferenças de temperaturas devem ser também menores. É esta relação que nos faz pensar em caracterizar os dados por meio de um modelo de Estatística Espacial.

Vamos analisar agora se esta tendência de variar conforme a localização acontece com estes dados.

É necessário, primeiramente, entender quão próximas estão estas estações, de forma mais intuitiva.

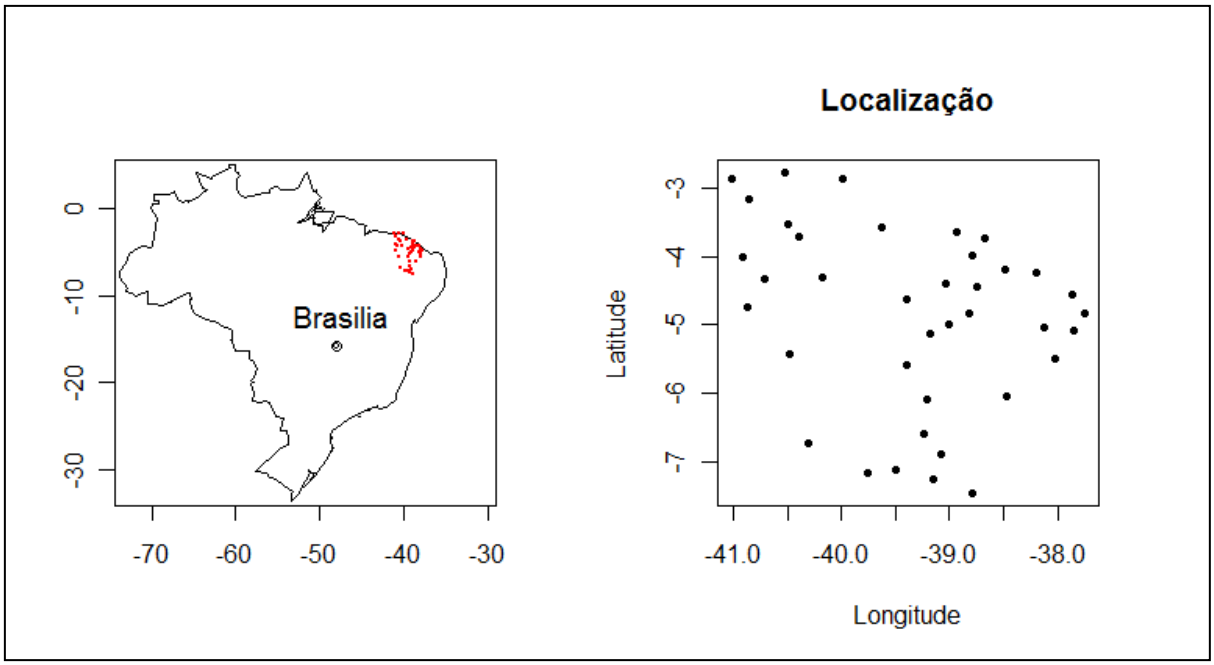



**Figura 7: Localização das estações no Brasil**

Verificamos que as estações estão localizadas no Nordeste do Brasil e, portanto, estão distribuídas em uma mesma região. Por exemplo, se alguma estação estivesse no Sul do Brasil, seria difícil comparar temperaturas e umidades desta estação com as demais do Nordeste. Já se as estações estivessem distribuídas em todo o país, tendo estações em todas as regiões, a comparação poderia ser feita. Daí a importância da análise da localização das estações. Assim, podemos perceber que não temos neste caso nenhum outlier, nenhum ponto com distância dos demais muito maior nem muito menor do que o comum.

As distâncias entre as estações variam de 0.2059126 a 5.0760713 graus.

Analisando o histograma da Figura 8, confirma-se que as distâncias são pequenas e pode-se perceber que as distâncias têm maior frequência entre 0,5 e 3,5.

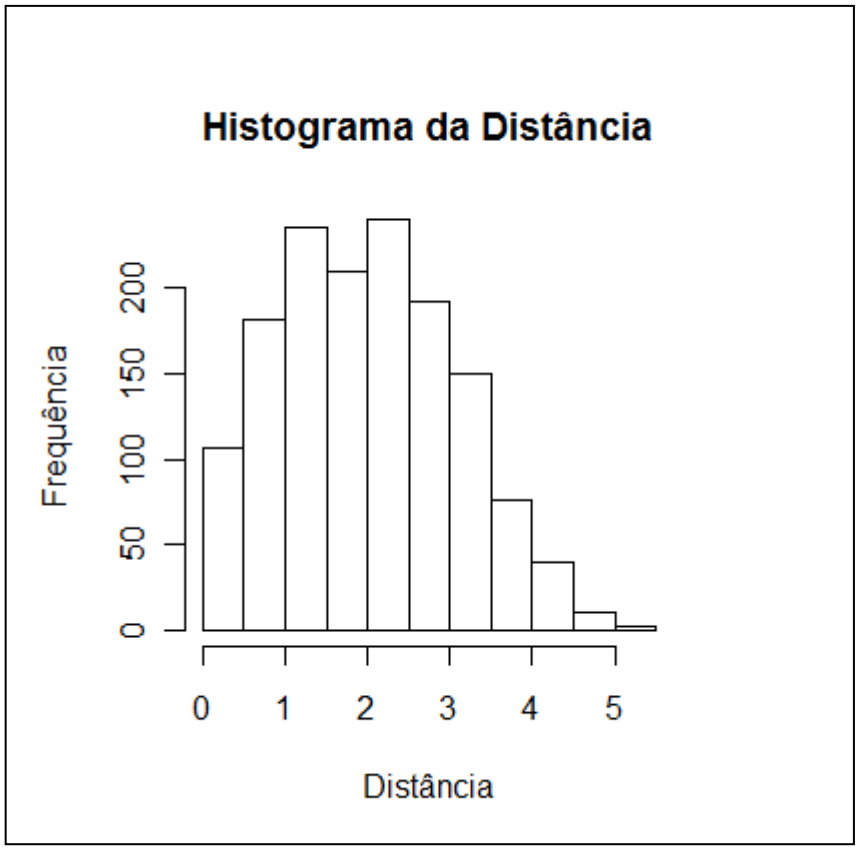



**Figura 8: Histograma das distâncias entre as estações**

Uma importante verificação é se as temperaturas e umidades se comportam de forma mais parecida em pontos mais próximos. Para isso, é importante a análise das Figuras 9 e 10.

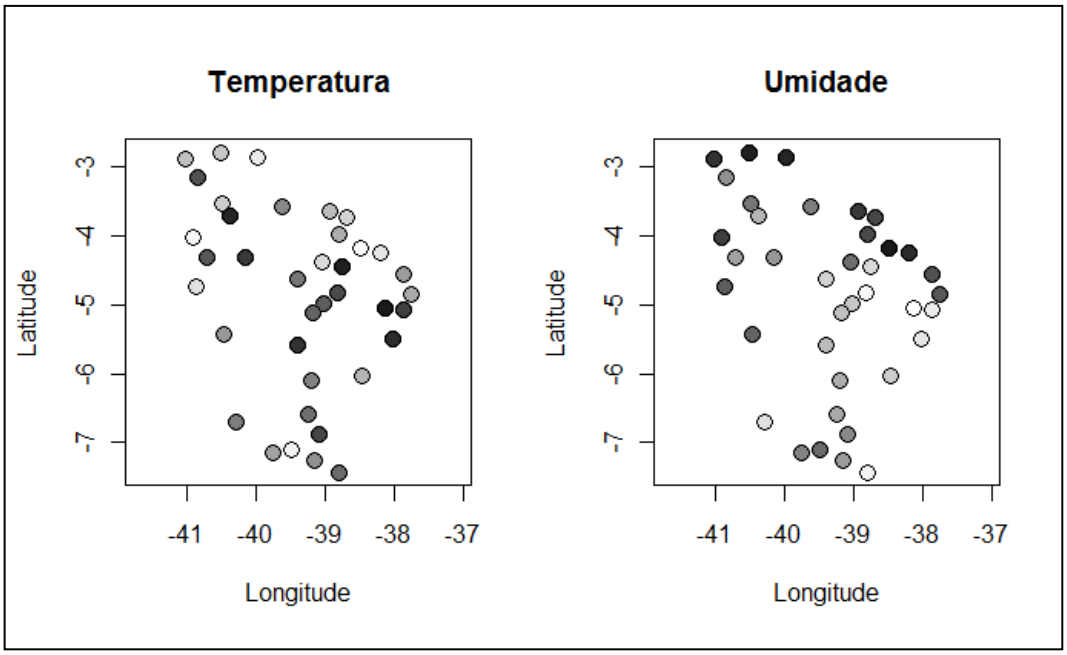



**Figura 9: Gráficos das estações com cores representando as medidas de temperatura e umidade**

As cores mais fortes significam valores maiores.

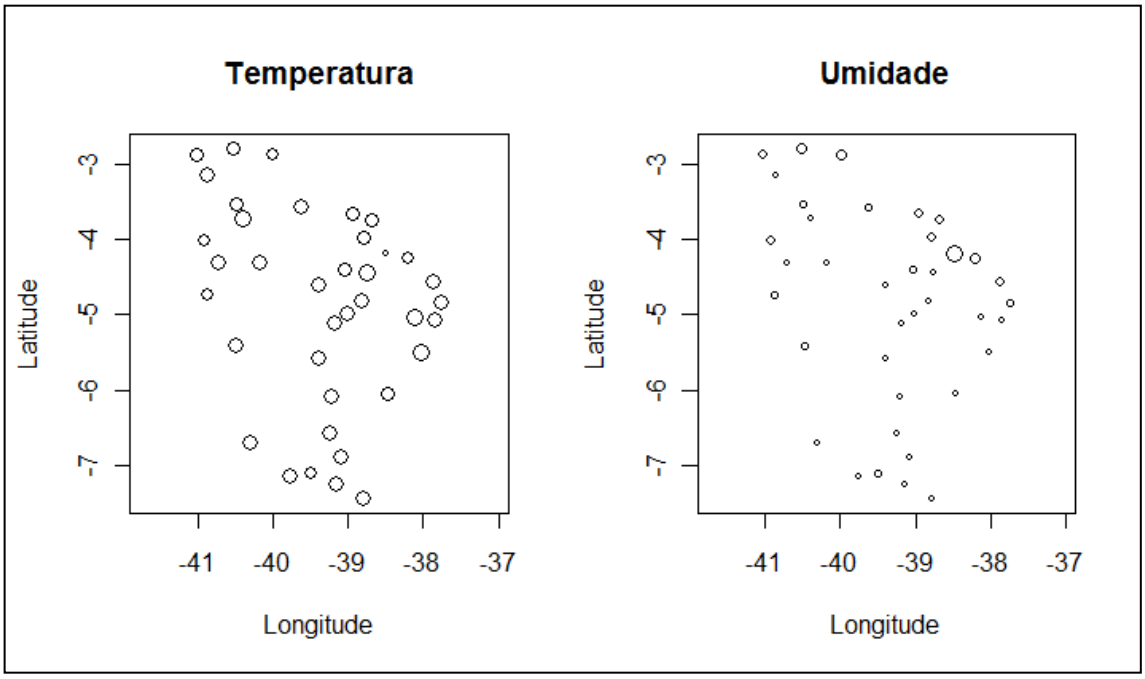



**Figura 10: Gráfico das estações com tamanhos representando as medidas de temperatura e umidade**

Os tamanhos maiores significam valores maiores.

É interessante perceber que no gráfico da temperatura onde os pontos estão mais escuros ou maiores, no gráfico da umidade estes pontos são os mais claros ou os menores, isto ocorre devido a relação inversa entre a umidade e a temperatura.

Também é importante analisarmos que, como são pontos na mesma região, temperaturas e umidades não deveriam ter diferenças extremas, e ao analisarmos os gráficos percebemos que os valores se comportam de forma homogênea, o que é mais perceptível nos dois últimos gráficos. As maiores variações nas medidas das umidades explicam-se pelo fato de no litoral a umidade ser maior.

Mencionamos variogramas no capítulo 4, e agora encontraremos os variogramas empíricos para a temperatura e para a umidade.

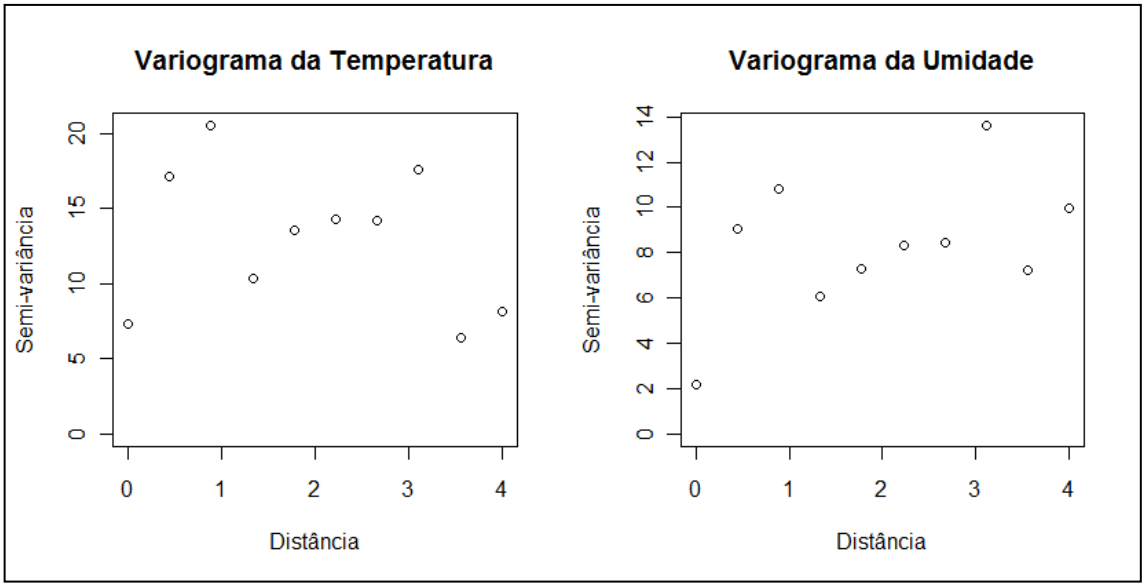



**Figura 11: Variogramas da Temperatura e da Umidade encontrados pelo Estimador Clássico**

No Estimador Módulo, no lugar de elevar a diferença ao quadrado, encontra-se o módulo da diferença.

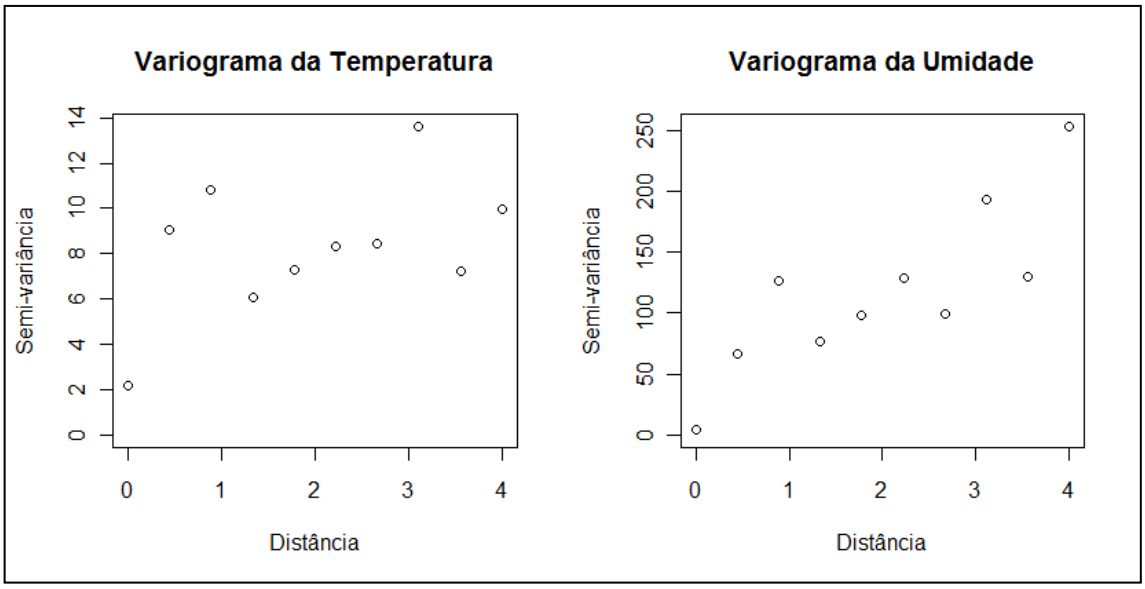



**Figura 12: Variogramas da Temperatura e da Umidade encontrados pelo Estimador Módulo**

Espera-se que a semi-variância aumente conforme a distância aumente. Não é possível ter um resultado ideal com estes dados porque existem poucos pontos, mas ainda assim é possível visualizar um aumento da semi-variância da temperatura e da umidade quando a distância aumenta. Este gráfico nos diz, portanto, que existem maiores variações da umidade e da temperatura quanto mais distantes são as localizações.

## 5.3 Inferência

Neste tópico, utilizando o software WinBUGS, modelaremos a temperatura por dois métodos: pela Regressão Linear e pela forma espacial. Os códigos utilizados estão descritos nos apêndices A1 e A2. Optaremos pelo modelo que melhor se adéqua aos dados através do Critério DIC (Deviance Information Criterion).

O DIC é definido da seguinte forma:

$$DIC = -2\,log\,p\left(y\middle|\tilde{\theta}\right) + 2p_D$$

onde:

$p_D = \overline{D} - D(\breve{\theta})$

$D(\theta) = -2\, ln\, p(y|\theta)$

$\breve{\theta} = E[\theta|y]$

$\overline{D} = E[D(\theta)|y]$

Quanto menor for o DIC, melhor é o ajuste do modelo.

### 5.3.1 Regressão

Nesta seção, será utilizada a abordagem de Regressão citada no capítulo 2, cuja variável resposta Y é a temperatura e a variável explicativa X é a umidade. Utilizamos uma cadeia de 30.000 iterações e as prioris especificadas foram:

$$\alpha \sim N(0\,, 0{,}001)$$

$$\beta \sim N(0\,, 0{,}001)$$

$$\tau^2 \sim Gamma(0{,}001\,, 0{,}001)$$

Os parâmetros utilizados no modelo devem convergir. Para verificação da convergência, analisaremos o comportamento da cadeia de cada parâmetro. O gráfico entre o valor do parâmetro e o número de iterações nos permite esta análise.

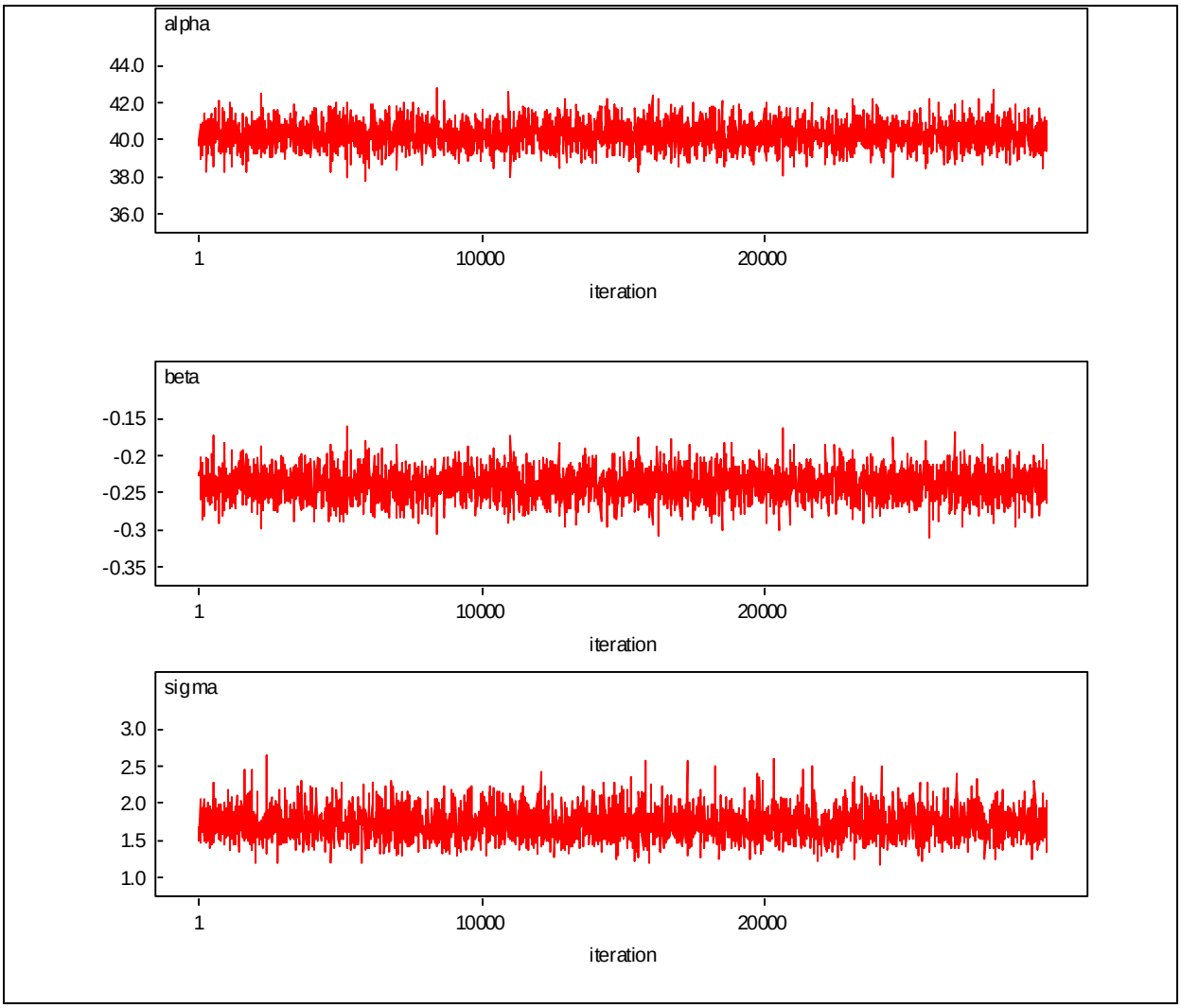



**Figura 13: Gráficos dos valores estimados dos parâmetros versus o número de iterações para verificação da convergência dos parâmetros do Modelo de Regressão**

Podemos perceber que os gráficos são aleatórios, ou seja, não apresentam tendências,vícios. Os parâmetros estão distribuídos aleatoriamente em torno de uma média, indicando a convergência dos parâmetros.

Para verificar a sensibilidade dos valores dos parâmetros, devemos analisar as seguintes estatísticas:

| Parâmetros | Média | Quantil 2,5% | Quantil 50% (mediana) | Quantil 97,5% |
|---|---|---|---|---|
| Alpha | 40,25 | 38,88 | 40,25 | 41,62 |
| Beta | -0,2385 | -0,278 | -0,2387 | -0,1983 |
| Sigma | 1,708 | 1,358 | 1,687 | 2,188 |

Ao analisarmos os gráficos das funções de densidade dos parâmetros alpha e beta na Figura 14, percebemos que os mesmos se aproximam da Distribuição Normal:

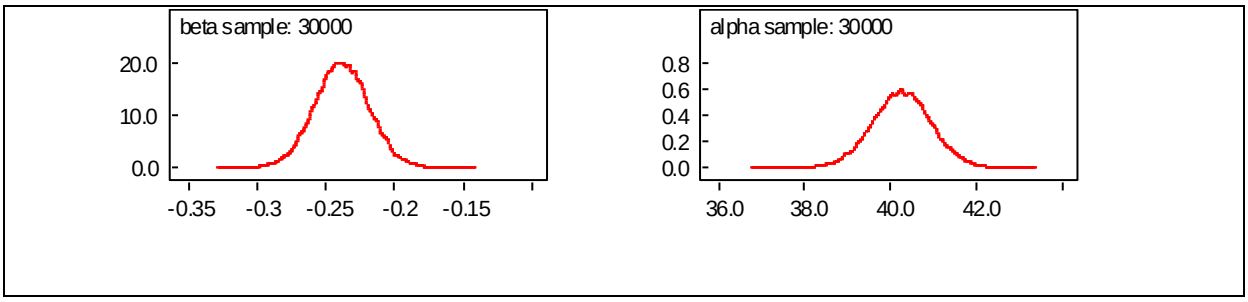



**Figura 14: Funções de Densidade dos parâmetros do Modelo de Regressão**

Na Figura 15, observamos os gráficos contendo os intervalos de confiança dos parâmetros utilizados no modelo.

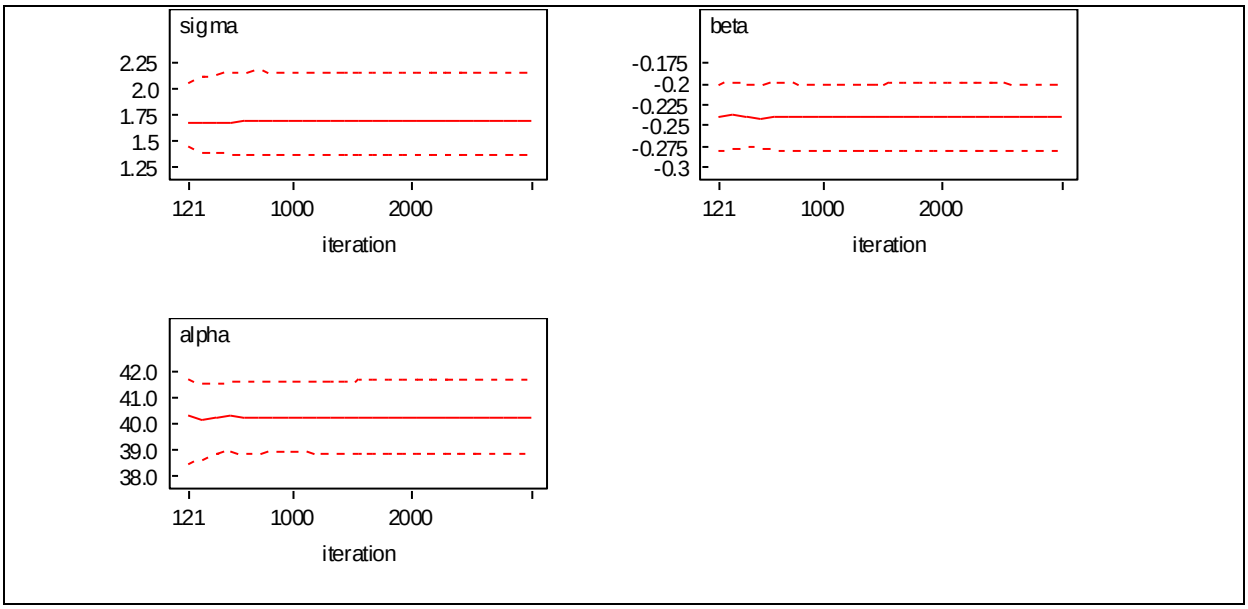



**Figura 15: Intervalos de Confiança dos parâmetros do Modelo de Regressão**

O DIC encontrado neste modelo foi 151,108.

### 5.3.2 Estatística Espacial

Nesta seção, utilizaremos a abordagem da Distribuição Condicional citada na subseção 4.4, cuja variável resposta Y é a temperatura e a variável explicativa X é a umidade. Foi utilizada uma cadeia de 30.000 iterações e as prioris especificadas foram:

$$\alpha \sim N(0\,,0{,}001)$$

$$\beta \sim N(0\,,0{,}001)$$

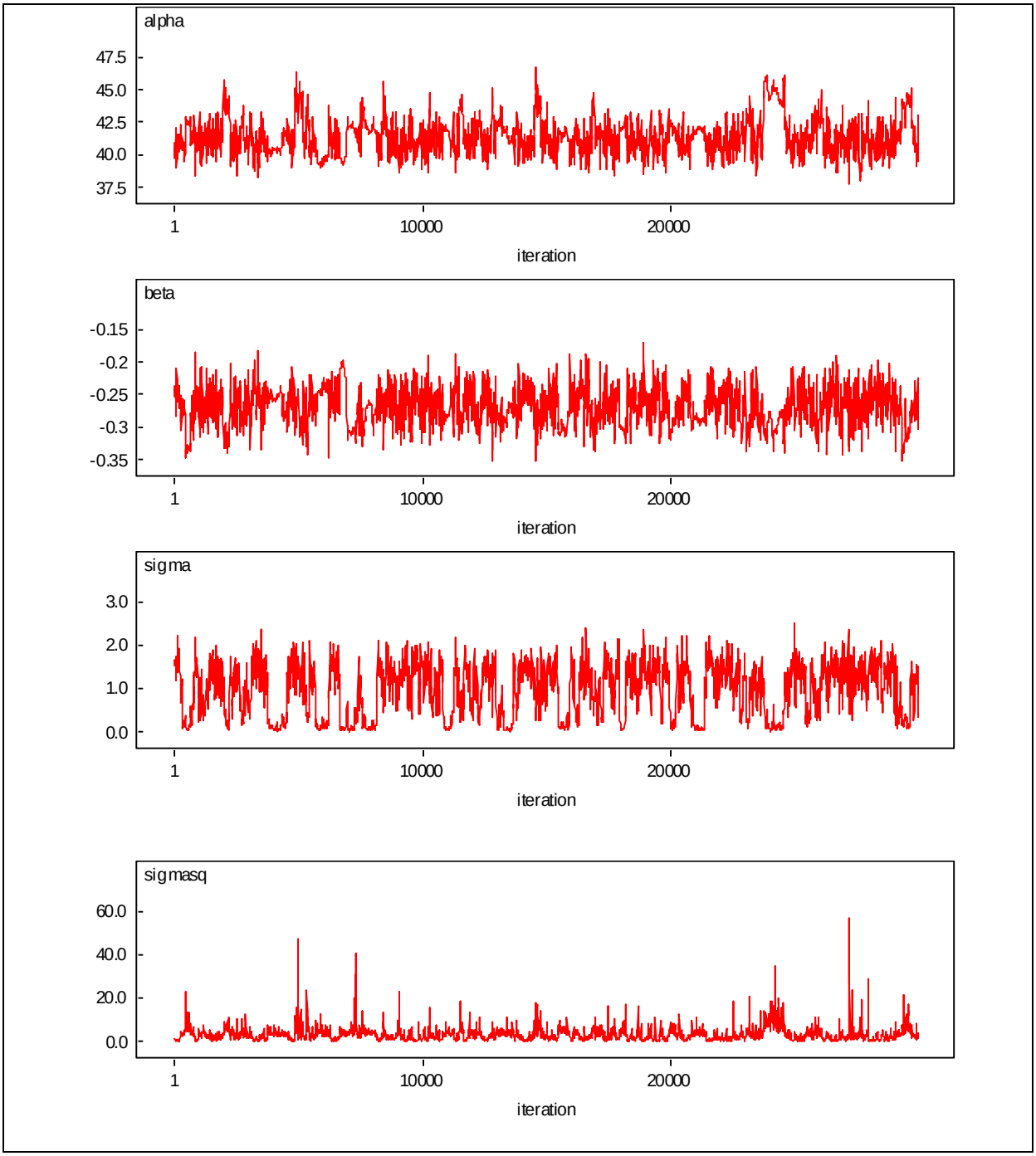
$$\tau^2 \sim Gamma(0{,}001\,,0{,}001)$$

$$spat.prec \sim Gamma(0{,}1\,,0{,}1)$$

$$\emptyset \sim Gamma(2\,,1)$$

Sendo $\tau^2$ o patamar parcial, spat.prec o efeito pepita e o $\emptyset$ o parâmetro de decaimento.

Os parâmetros utilizados no modelo devem convergir. Para verificação da convergência, analisaremos o comportamento da cadeia de cada parâmetro. O gráfico entre o valor do parâmetro e o número de iterações nos permite esta análise.



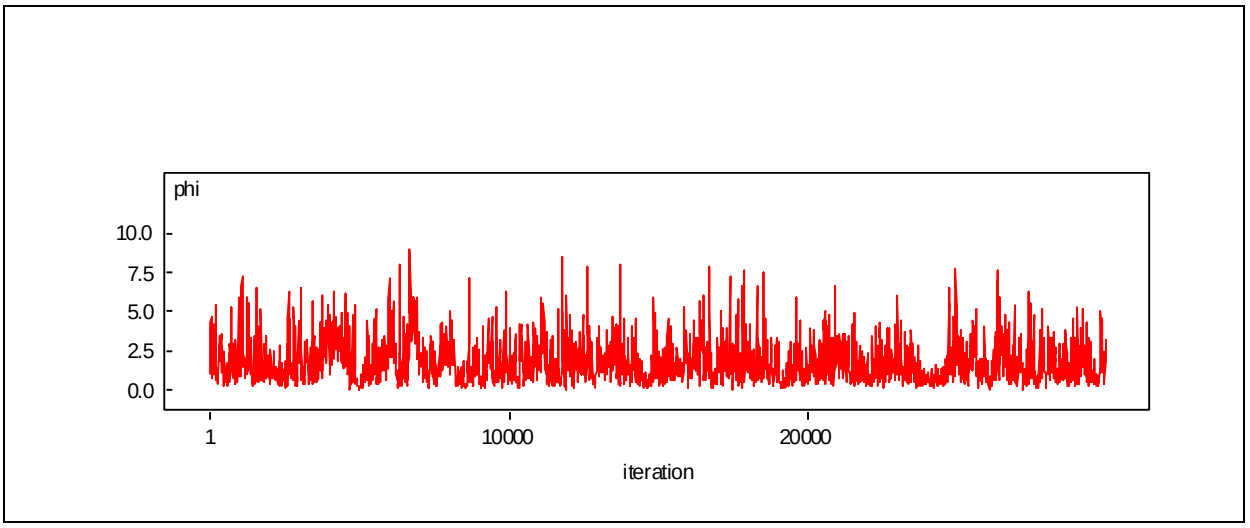



**Figura 16: Gráficos dos valores estimados dos parâmetros versus o número de iterações para verificação da convergência dos parâmetros do Modelo Espacial**

Podemos perceber que os gráficos são aleatórios, ou seja, não apresentam tendências,vícios. Os parâmetros estão distribuídos aleatoriamente em torno de uma média, indicando a convergência dos parâmetros.

Para a sensibilidade dos valores dos parâmetros, analisaremos as seguintes estatísticas:

| **Parâmetros** | **Média** | **Quantil 2,5%** | **Quantil 50% (mediana)** | **Quantil 97,5%** |
|---|---|---|---|---|
| Alpha | 27,34 | 23,83 | 26,73 | 33,07 |
| Beta | -0,2574 | -0,4297 | -0,2368 | -0,1532 |
| Phi | 0,001303 | 1.28E-5 | 7.826E-4 | 0.005464 |
| Sigma | 0,05901 | 0.02077 | 0.05085 | 0.1421 |
| Sigmasq | 26,65 | 6.219 | 15.95 | 127.4 |

Podemos perceber que o parâmetro beta é significativo e negativo, o que demonstra mais uma vez que a temperatura é inversamente proporcional à umidade tanto para este modelo, quanto para o modelo de Regressão.

Ao analisarmos os gráficos das funções de densidade dos parâmetros alpha e beta na Figura 17, percebemos que os mesmos se aproximam da Distribuição Normal.

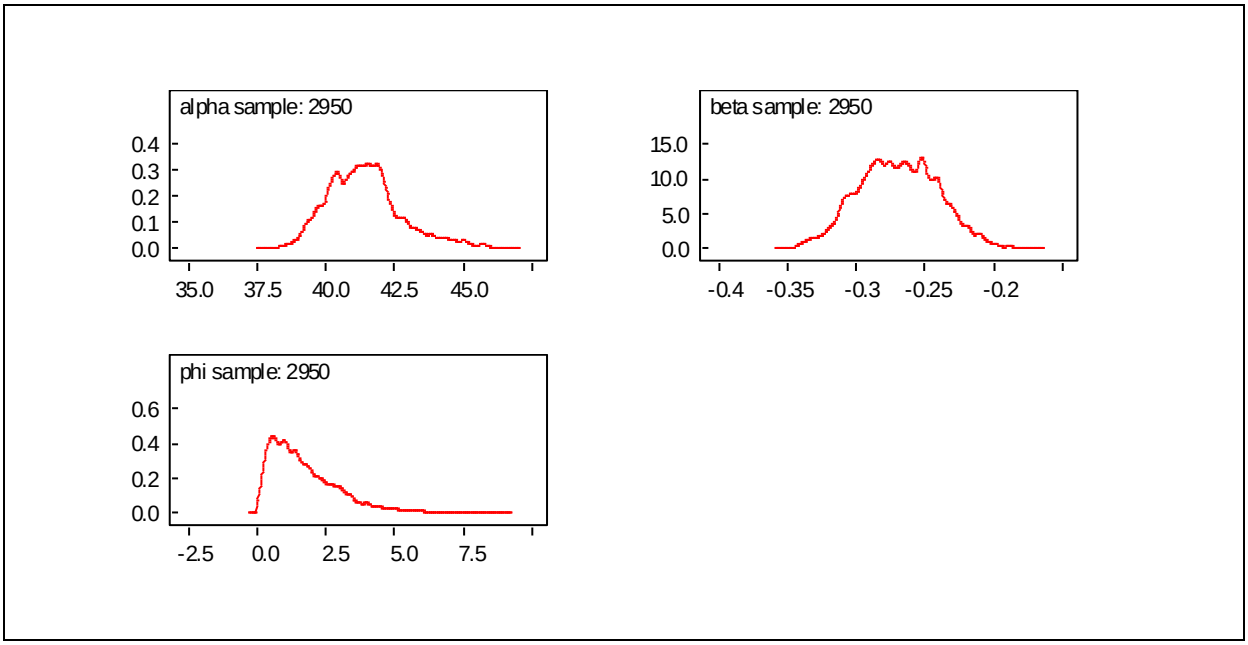



**Figura 17: Funções de Densidade dos parâmetros do Modelo Espacial**

Na Figura 18, observamos os gráficos contendo os intervalos de confiança dos parâmetros utilizados no modelo.

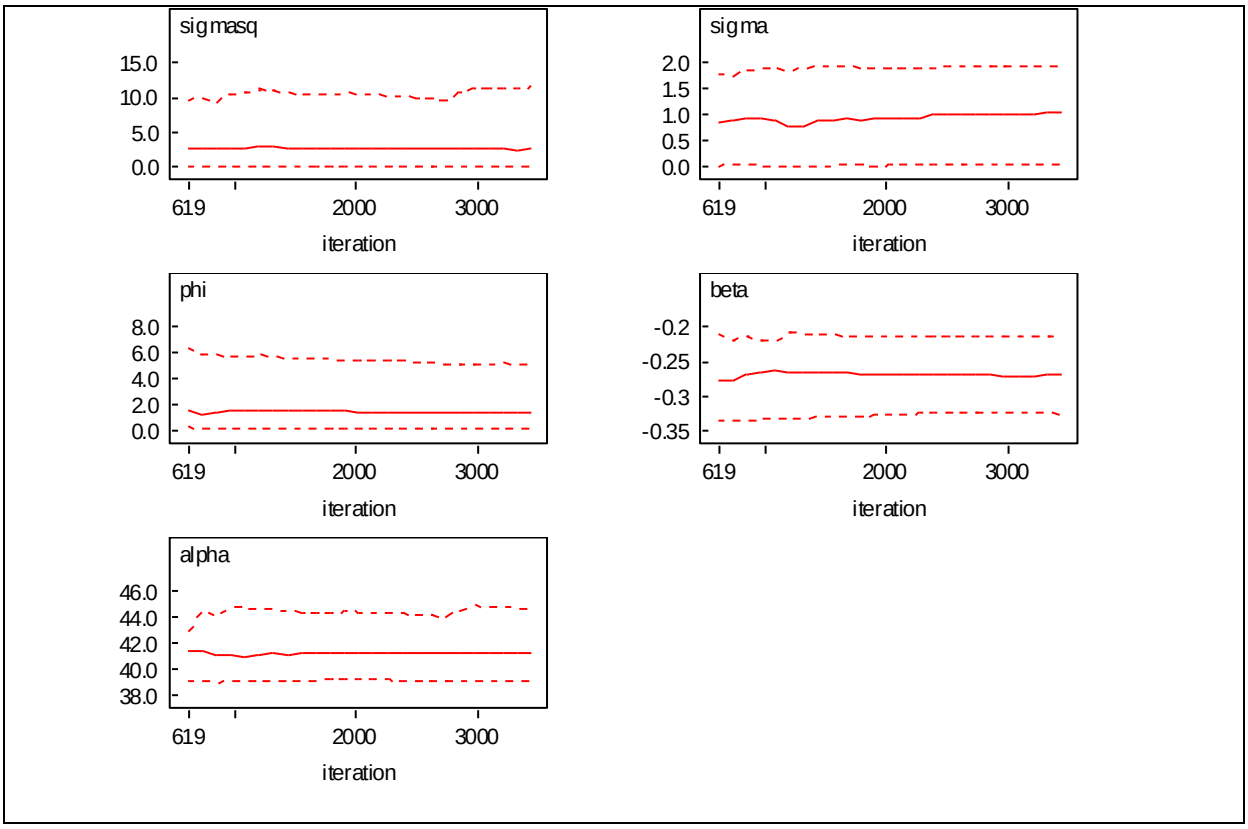



**Figura 18: Intervalos de Confiança dos parâmetros do Modelo Espacial**

O DIC encontrado neste modelo foi -941,303, valor menor do que o encontrado no modelo de Regressão. Isto demonstra que o modelo que melhor se ajusta é o modelo espacial.

# 6 Conclusão

A partir dos estudos de Análise de Regressão e Processos Estocásticos, foi possível entender o procedimento da Estatística Espacial. Para os dados que utilizamos de temperaturas e umidades, verificamos que dentre as três divisões da Estatística Espacial, a Geoestatística é a que nos permite adequação dos dados.

Ao analisarmos os dados, verificamos que é verídica a dependência espacial, e que a temperatura e a umidade são inversamente proporcionais. Foi possível a

sensibilidade da dependência espacial quando verificamos graficamente valores mais próximos para estações com menores distâncias.

Os dados foram modelados pela Regressão Simples e pela forma espacial para que verificássemos o modelo que melhor descrevesse a relação existente entre as variáveis. Inicialmente já percebemos a dificuldade de modelar os dados pela Regressão Simples, já que os mesmos não satisfizeram duas condições necessárias para a utilização desta técnica: os resíduos não são independentes e a variância dos resíduos não é constante.

Foi necessária a utilização dos métodos MCMC, já que não existia forma fechada para as distribuições a posteriori dos parâmetros. Implementamos estes modelos pelo WinBUGS e obtivemos resultados esperados: convergência dos parâmetros, comportamento do parâmetro Beta seguindo uma Distribuição Normal e DIC menor para o modelo espacial. Com estes resultados, podemos concluir pela melhor adequação do modelo espacial aos dados de temperatura e umidade relativa do ar.

## Apêndice

### A1 Código WinBUGS – Modelo de Regressão Simples

```
model
    {
        for( i in 1 : N ) {
                Y[i] ~ dnorm(mu[i],tau.c)
                mu[i] <- alpha + beta * x[i]
        }
        alpha ~ dnorm(0,0.001)
        beta ~ dnorm(0,0.001)
        tau.c ~ dgamma(0.001,0.001)
        sigma <- 1 / sqrt(tau.c)
    }

Data
list(N=38, Y = c(29.1, 34.0, 34.4, 33.3, 36.0, 30.1, 29.1, 30.9, 34.5, 30.3, 34.6, 28.3,
35.3, 35.3, 33.5, 35.2, 33.7, 32.9, 30.7, 35.5, 32.6, 34.9, 30.5, 33.6, 15.8, 32.1, 29.5,
35.0, 34.9, 35.5, 36.6, 35.5, 33.0, 27.5, 31.0, 35.5, 35.9, 35.8), x = c(54.4, 24.0, 22.4,
38.4, 22.4, 33.6, 52.8, 46.4, 22.4, 41.6, 24.0, 33.6, 25.6, 19.2, 33.6, 24.0, 28.8, 35.2,
54.4, 25.6, 38.4, 19.2, 32.0, 24.0, 89.6, 22.4, 33.6, 22.4, 22.4, 20.8, 17.6, 24.0, 25.6,
41.6, 44.8, 22.4, 22.4, 20.8))

Inits1 list(alpha = 0, beta = 0, tau.c = 1)
```

### A2 Código WinBUGS – Modelo Espacial

```
model
 {
 for (i in 1:N) {
 Y[i] ~dnorm(mu [i], tau.c)
```

```
mu[i] <- alpha + beta * x[i] + W[i]
muW[i] <- 0
}

alpha ~ dnorm(0,0.001)
beta ~ dnorm(0,0.001)
tau.c ~ dgamma(0.001,0.001)
sigma <- 1 / sqrt(tau.c)

W[1:N] ~ spatial.exp(muW[], local1[], local2[], spat.prec, phi, 1)
phi~dgamma(2, 1)
spat.prec ~dgamma(0.10, 0.10)
sigmasq <- 1/spat.prec

}
```

Data

```
list(N=38, Y = c(29.1, 34.0, 34.4, 33.3, 36.0, 30.1, 29.1, 30.9, 34.5, 30.3, 34.6, 28.3,
35.3, 35.3, 33.5, 35.2, 33.7, 32.9, 30.7, 35.5, 32.6, 34.9, 30.5, 33.6, 15.8,  32.1, 29.5,
35.0, 34.9, 35.5, 36.6, 35.5, 33.0, 27.5, 31.0, 35.5, 35.9, 35.8  ), x = c(54.4, 24.0,
22.4, 38.4, 22.4, 33.6, 52.8, 46.4, 22.4, 41.6, 24.0,  33.6, 25.6, 19.2, 33.6, 24.0, 28.8,
35.2, 54.4, 25.6, 38.4, 19.2, 32.0, 24.0, 89.6, 22.4, 33.6, 22.4, 22.4, 20.8, 17.6, 24.0,
25.6, 41.6, 44.8, 22.4, 22.4, 20.8),local1 = c(-39.99,-39.21,-40.3,-37.87,-38.75,-
39.04,-38.2,-41.02,-39.39,-38.68,-39.24,-39.5,-40.86,-38.82,-40.48,-40.71,-39.62,-
37.75,-40.52,-39.08,-38.79,-38.79,-40.49,-39.15,-38.49,-38.47,-40.87,-39.01,-39.18,-
37.85,-38.12,-40.17,-39.76,-40.91,-38.94,-39.39,-40.39,-38.02), local2 = c(-2.87,-
6.09,-6.71,-4.56,-4.44,-4.39,-4.24,-2.88,-4.62,-3.74,-6.58,-7.1,-3.16,-4.82,-5.42,-
4.32,-3.57,-4.84,-2.79,-6.89,-3.98,-7.44,-3.54,-7.25,-4.19,-6.04,-4.74,-4.99,-5.12,-
5.08,-5.04,-4.31,-7.15,-4.02,-3.65,-5.58,-3.72,-5.49))
```

```
Inits1 list(alpha = 40, beta = -0.24, tau.c = 0.5, phi = 1, spat.prec =
0.5,W=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

## Referências Bibliográficas


- Casella, George e George, Edward I., 1992, *Explaining the Gibbs Sampler*, volume 46, The American Statistician
- Draper, Normand Richard, 1998, *Applied Regression Analysis*, 3ª edição, New York : J. Wiley & Sons
- Druck, S.; Carvalho, M.S.; Câmara, G.; Monteiro, A.V.M. (eds) *"Análise Espacial de Dados Geográficos"*. Brasília, EMBRAPA, 2004
- Gamerman, D. (1997) *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*. Chapman & Hall.
- Gelfand, Alan E. et al, 2003, *Hierarchical modeling and analysis for spatial data*, Chapman & Hall/CRC
- Marques, Mauro S. de F.,1996, *Elementos de Processos Estocásticos: minicurso*, IMECC/UNICAMP
- Miller, Kenneth S.,1962-64,*Multidimensional Gaussian Distributions*, John Wiley and Sons, Inc.
- Morettin, Pedro A e Toloi, Clélia M.C., 2004, *Análise de Séries Temporais*, Editora Edgar Blucher
- Ribeiro Jr, P.J. and Diggle, P.J. (2001) geoR: A Package for Geostatistical Analysis.
- Ross, Sheldon M.,1997,*Introduction to Probability Models*,6ª edição, Academic Press
- Spiegelhalter, D.J., Thomas, A. and Best, N.G. (2002) WinBugs Version 1.4 User Manual. Technical Report. Cambridge: Medical Research Council Biostatistics.