UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

MODELAGEM GEOESTATÍSTICA DA DISTRIBUIÇÃO ESPACIAL DA GASOMETRIA DE SUPERFÍCIE NA ATIVIDADE DE E&P DE PETRÓLEO E GÁS

DANIEL PROVENZI

Orientador: Professor Dr. Fernando Hepp Pulgati

Porto Alegre

2012

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

MODELAGEM GEOESTATÍSTICA DA DISTRIBUIÇÃO ESPACIAL DA GASOMETRIA DE SUPERFÍCIE NA ATIVIDADE DE E&P DE PETRÓLEO E GÁS

DANIEL PROVENZI

Monografia apresentada para a Universidade Federal do Rio Grande do Sul para a obtenção do grau de Bacharel em Estatística, sob orientação do Prof. Dr. Fernando Hepp Pulgati

Banca Examinadora:
Prof. Dr. Fernando Hepp Pulgati
Dieter Schwanke (Me.em Ciências da Computação - PPGC/ UFRGS)

Porto Alegre
2012

AGRADECIMENTOS

Quero agradecer ao professor e orientador Fernando Pulgati pelos ensinamentos, paciência e apoio nesta Monografia, à professora Jandyra Fachel pelo incentivo e apoio oferecido junto ao NAE e aos demais professores aos quais fui aluno pelos ensinamentos transmitidos.

Agradeço a toda minha família por caminharem sempre junto comigo em todos os momentos. Agradecer a Ali, minha namorada, por todo amor e compreensão durante esta investida.

RESUMO

O estudo da distribuição espacial da gasometria de superfície vem sendo algo cada vez mais presente nos meios de pesquisa. Relacionado a isso, está o interesse global em encontrar métodos que contribuam com a redução dos custos envolvidos na prospecção de petróleo e gás. A grande variabilidade e os fatores confundidores associados às medições dos gases em superfície dificultam a caracterização da fonte geradora do fenômeno. No presente estudo, a amostra é formada por dados geoquímicos e caracterizadores das condições de uso e cobertura do solo. Analisar a distribuição destes gases de forma georeferenciada é essencial para entender o comportamento do processo, uma vez que amostras mais próximas se assemelham mais em relação às mais distantes. Neste trabalho, a modelagem da dependência espacial dos gases metano, etano e eteno faz uso de métodos geoestatísticos através de estimadores de máxima verossimilhança e de pressuposições gaussianas no ajuste de modelos que permitam verificar a associação de fatores físicos que podem influenciar na distribuição do fenômeno. Este trabalho objetiva apresentar através da produção de mapas de probabilidade, uma ferramenta probabilística que, juntamente com informações que caracterizam uma maior favorabilidade à presença de hidrocarbonetos de fonte termogênica, auxilie geograficamente na determinação áreas que demonstram um cenário favorável à identificação de regiões potencialmente exploratórias.

ABSTRACT

Studies about Surface Gasometry's Space Distribution have been becoming a frequent topic amongst the research community. Additionally, there is an increasing global concern on finding better cost saving methods for oil and gas prospecting. A great variety of misleading factors related to gas measurements on surfaces set challenging hurdles on identifying the origin of such phenomenon. In this study, samples are set by geochemical and geophysical data that were observed and collected in the region. Geo-referential gas distribution analysis is crucial to understand the behavior of such process. In this research, geostatistics methods, through maximum likelihood estimators and Gaussian assumptions, were used to modeling the spatial dependence of methane, ethane and ethene. Such method allowed verifying physical factors that might influence over gas' spatial distribution. The findings are presented in probability maps, a statistic tool that, jointly with the information of thermal hydrocarbon presence likelihood, helps to geographically identifying areas that holds favorable conditions for potentially exploratory regions.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O setor da energia e de recursos naturais é fundamental para a economia mundial. O crescente aumento na demanda de derivados do petróleo é um dos maiores desafios para os atuantes neste setor de atividade. Neste cenário, petroleiras investem pesado nas atividades de exploração e produção de petróleo (E&P), em especial na prospecção de novas reservas.

Ao passar das décadas de exploração, percebeu-se que para encontrar jazidas de hidrocarbonetos de volume relevante era eminente que um número de requisitos geológicos ocorresse simultaneamente nas bacias sedimentares. Estudar estas características de maneira integrada buscando detectar as condições mais favoráveis diminuindo o risco exploratório envolvido nas perfurações de poços que são de elevado custo.

1.1 CONTEXTO

Na prospecção de petróleo e gás a primeira atividade de reconhecimento da área a ser estudada envolve a avaliação da distribuição das concentrações de gases na superfície da região em estudo. Esta atividade de reconhecimento e exploração é conhecida como gasometria de superfície. Ela inicia com o estudo da água e do solo da região onde estão localizados os focos de gás. Dentre os problemas encontrados na análise destes dados destaca-se a investigação da origem do gás, podendo esta ser biogênica ou termogênica.

Considera-se de origem biogênica os gases que se formam a partir de substâncias orgânicas procedentes da superfície terrestre, como os detritos orgânicos. Os de origem termogênica são provenientes de uma fonte em sub-superfície e constituem evidências da presença de reservatórios de óleo e/ou gás.

Neste contexto, busca-se através deste estudo desenvolver uma metodologia que auxilie em uma melhor caracterização e reconhecimento das acumulações de hidrocarbonetos na região analisada.

Este trabalho tem enfoque na utilização de técnicas geoestatísticas em dados provenientes da geoquímica de superfície (gasometria) da região em estudo, que por motivos de confidencialidade da informação, tem sua identidade preservada e seu sistema de coordenadas reposicionado.

A geoestatística é caracterizada por estudar fenômenos que envolvem processos espaciais indexados sobre o espaço contínuo. A geoestatística, segundo Cressie (1993), é uma das três áreas da Estatística Espacial. Sua modelagem trata da estimação e predição, sendo o primeiro relacionado à inferência dos parâmetros de um modelo que expresse a dependência espacial, e o segundo, em predizer valores não observados na região em estudo com base neste modelo.

Portanto, desenvolver um modelo que isole os fatores ambientais e geológicos que podem influenciar diretamente nos gases medidos na superfície é de grande importância para distinguir reais evidências da possibilidade da presença de reservatórios de óleo e/ou gás em sub-superfície.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O presente estudo propõe analisar a distribuição espacial da gasometria de superfície das variáveis: Metano, Etano e Eteno, ajustando importantes fatores físicos associados às fontes de origem biogênicas (Umidade, Tipo de Solo, Cor e Uso do Solo).

1.2.2 Objetivos Específicos

Pretende-se produzir mapas de probabilidade para os diferentes elementos medidos, integrando-os de forma a apresentar uma metodologia que permita construir cenários favoráveis à identificação de regiões com indícios de hidrocarbonetos de origem termogênica sabidamente associados à presença de óleo e gás.

Oferecer através deste documento um material de apoio ao uso das técnicas de geoestatística através do pacote geoR do software livre R, demonstrando a utilidade desta técnica estatística de uma maneira sucinta e de fácil leitura.

1.3 ESTRUTURA DO TRABALHO

A proposta da apresentação deste trabalho é trazer uma breve revisão literária sobre a técnica geoestatística, pois a mesma já é vastamente difundida na literatura acadêmica, dando ênfase à criação de um registro de fácil compreensão, com uma abordagem prática do método.

Para isso, é apresentado no capítulo 2 o referencial teórico a respeito da técnica Geoestatística. A metodologia adotada é apresentada no capítulo 3 e, por fim, os resultados das análises e dos mapas de probabilidades, além da conclusão estão dispostos nos capítulos 4 e 5, respectivamente.

2 REFERENCIAL TEÓRICO

2.1 GEOESTATÍSTICA

Geoestatística é sub-ramo de estatística espacial onde os dados consistem de uma amostra finita de valores medidos que se referem a um fenômeno subjacente espacialmente contínuo (DIGGLE; RIBEIRO, 2007).

O termo geoestatística é utilizado para identificar uma parte dos métodos de estatística espacial, na qual o modelo utilizado descreve uma variação contínua das observações no espaço. A geoestatística modela a distribuição espacial associando o grau de dependência a medidas de distância e de direção entre os pontos amostrados (VIOLA, 2007; ODA-SOUZA, 2009).

2.1.1 História e contexto

Criada no final da década de cinquenta, a geoestatística surgiu da necessidade de modelagem de recursos geológicos, como por exemplo, a caracterização da concentração de metais em jazidas minerais ou o estudo da qualidade de águas subterrâneas. Daí surgiu o prefixo *geo*, que foi utilizado pela primeira vez por Hart (1954). Posteriormente, Georges Matheron contribuiu decisivamente para a atual orientação com a publicação de alguns artigos onde desenvolveu o conceito de *variável regionalizada* (MATHERON, 1962; MATHERON, 1963; MATHERON, 1971).

Atualmente, a geoestatística é utilizada por diversas áreas da ciência, tais como ciências da terra, do ambiente e até mesmo da saúde. O interesse que estes diversos ramos encontram nos métodos geoestatísticos resultam da sua capacidade única para modelar um processo estocástico observado em localizações fixas e que se pretende estudá-lo numa região com índice espacial contínuo.

Duas fases no trabalho da geoestatística são fundamentais, a obtenção do variograma e a *krigagem*. O variograma é o instrumento que permite caracterizar a estrutura de dependência existente nas observações do processo. É um instrumento valioso, pois vai influenciar decisivamente a fase de krigagem. A designação de krigagem foi introduzida por G. Matheron, em honra a D. G. Krige, para designar um conjunto de métodos de predição de futuras observações do processo.

Existe na literatura acadêmica uma vasta quantidade de trabalhos que descrevem detalhadamente toda a fundamentação matemática das técnicas geoestatísticas e os conceitos envolvidos na sua utilização. Neste trabalho, optou-se por descrever seu conteúdo com ênfase na metodologia de aplicação das técnicas geoestatísticas, deixando como legado um registro em português do uso da ferramenta *geoR* como instrumento na aplicação destas técnicas. Tendo isso em mente, será apresentada apenas uma breve revisão dos conceitos envolvidos durante uma análise geoestatística baseada em modelos.

A expressão geoestatística baseada em modelos foi trazida por Diggle, Tawn e Moyeed (1998). Nela, os problemas geoestatísticos são fundamentados na aplicação de métodos estatísticos formais, com a explicitação de um modelo e de métodos de inferência baseados na máxima verossimilhança.

2.1.2 Terminologia e notação

A notação básica utilizada para dados geoestatísticos univariados é (DIGGLE; RIBEIRO, 2007),

$$(x_i, y_i): i = 1, \cdots, n,$$

onde x_i identifica uma localização espacial (tipicamente no espaço bidimensional) e y_i é um valor escalar, da chamada variável resposta, associado à localização $x_i$sobre uma região em estudo A. Será assumido que o plano amostral para os locais x_i ou é determinista (por exemplo, o x_i podem formar uma grade sobre a região de estudo), ou estocasticamente independente do processo que gera as medições y_i. Cada y_i é uma realização de uma variável aleatória Y_i cuja distribuição é dependente do valor do local de um x_i de um processo estocástico espacialmente contínuo subjacente denominado S (x), que não é diretamente observável.

Portanto, a forma básica de um modelo geoestatístico incorpora pelo menos dois elementos: um processo estocástico de valor real $\{S\,(x): x \in A\}$, que é comumente considerado uma realização parcial de um processo estocástico $\{S\,(x): x \in \mathbb{R}^2\}$ em todo o plano, e uma distribuição multivariada para a variável aleatória $Y = (Y_1, \ldots, Y_n)$ condicionada à $S\,(\cdot)$.

2.1.3 Condições de estacionariedade

Para garantir determinadas propriedades fundamentais, condições de estacionariedade em todo seu domínio são desejáveis nos processos estocásticos. A regularidade nas variáveis aleatórias é garantida por estas condições, possibilitando assim, a posterior utilização da metodologia de inferência estatística. Hilário (2009) define as três condições de estacionariedade a seguir:

2.1.3.1 Estacionariedade forte ou estrita

Um processo é dito estacionariamente forte quando este se mantém invariante sempre que aplicada uma translação a qualquer um dos conjuntos de suas localizações. Deste modo a função densidade de probabilidade e é independente do ponto do domínio onde estas localizações se encontram. Isto implica que as variáveis aleatórias do processo sejam identicamente distribuídas.

Esta condição de estacionariedade é bastante restritiva para ser assumida em diversos fenômenos naturais, Por isso, outros conceitos menos restritivos para a estacionariedade são adotados.

2.1.3.2 Estacionariedade fraca ou de segunda ordem

É definido como estacionário de segunda-ordem, ou estacionária fraco, o processo que possui sua média constante (2.1) para todo o espaço contínuo $\mathcal{D}$ independente da variação da sua localização espacial. Além disso, C, que é a *função de covariância estacionária* apresentada em (2.2), conhecida como *covariograma* do processo, garante que a covariância entre duas quaisquer variáveis aleatórias do processo dependam apenas da diferença entre as suas localizações.

$$E[Z(s)] = \mu(s) = \mu \in \mathbb{R}\,; \quad \forall_{s_i, s_j \in D} \tag{2.1}$$

$$Cov[Z(s_i), Z(s_j)] = C(s_i - s_j)\,; \quad \forall_{s_i, s_j \in D} \tag{2.2}$$

2.1.3.3 Estacionariedade intrínseca

É denominada estacionariedade intrínseca quando de um processo tem-se

$$E[Z(s_i) - Z(s_j)] = 0 \; ; \; \forall_{s_i, s_j \in D} \quad (2.3)$$

$$Var[Z(s_i) - Z(s_j)] = 2\gamma(s_i - s_j) \; ; \; \forall_{s_i, s_j \in D} \quad (2.4)$$

onde as variáveis aleatórias $Z(s_i) - Z(s_j)$ são denominadas incrementos do processo. A condição (2.3) refere-se ao fato de a estacionariedade média estar expressa em termos dos incrementos. Já a condição (2.4) garante que a variância de um incremento apenas depende do vetor da diferença entre as localizações i e j, onde a função $2\gamma()$ é conhecida como variograma, que é um parâmetro fundamental da geoestatística, sendo a principal ferramenta para todo o processo de estimação e predição dos dados.

2.1.4 Variograma

O variograma é uma função fundamental na geoestatística, uma vez que modela a estrutura de dependência do processo. Nele é traduzida a variância dos incrementos do processo.

Considerando $h = \|s_i - s_j\|$ como o vetor de distâncias que separa dois pontos, a função variograma é definida como (MIRANDA, 2009).

$$2\gamma(h) = Var[Z(s) - Z(s + h)] \quad (2.5)$$

O variograma é dito *isotrópico* se depender apenas da norma do vector h. Na grande maioria dos casos o variograma é uma função crescente. A função $\gamma(h)$ representa o semivariograma, que é metade do variograma. Cressie (1993) apresenta as propriedades que caracterizam um *variograma válido* e afirma que qualquer função que seja um variograma válido pode ser utilizada como variograma de algum processo geoestatístico.

Os parâmetros do semivariograma são descritos em (DRUCK et al., 2004)

2.1.4.1 Patamar

À medida que se distanciam espacialmente as observações, o valor do semivariograma tende a um valor constante. Este valor recebe o nome de patamar. O patamar é uma característica exclusiva de variogramas de processos

estacionários de segunda ordem. A partir deste ponto a dependência espacial é desconsiderada, pois a variância da diferença entre os pares tornam-se aproximadamente constantes.

2.1.4.2 Alcance

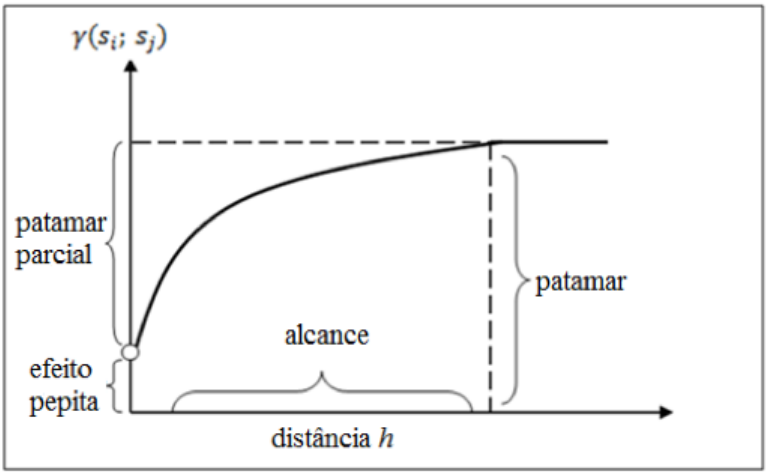
Alcance é a distância em que, a partir dela, a correlação existente entre as variáveis aleatórias é não significativa, por isso as variáveis aleatórias que estão distanciadas acima do alcance, podem ser consideradas não correlacionadas.

Consequentemente, alcance e patamar são conceitos dependentes. Apenas pode haver alcance se houver patamar. Com isso, o alcance só existe em variogramas de processos estacionários de segunda ordem.

2.1.4.3 Efeito Pepita

O variograma frequentemente apresenta uma descontinuidade na origem. O termo efeito pepita refere-se a esta descontinuidade. Esta descontinuidade pode ser originada por uma variação em microescala do processo que o variograma não está conseguindo modelar. Ainda, a descontinuidade também pode ser devida ao erro presente em cada medição, onde duas observações em uma mesma localização poderem ser bastante distintas.

Figura 1 - Parâmetros do semivariograma



Fonte: Cressie (1993) alterada pelo autor(2012)

2.1.4.4 Modelos de variograma isotrópico

Diversos modelos paramétricos são propostos na literatura para representar variogramas isotrópicos. Estes variogramas são agrupados por famílias, onde estas representam o conjunto de variogramas definidos pela mesma expressão algébrica. Em Journel e Huijbregts (1978) são apresentadas informações detalhadas sobre modelos paramétricos de variogramas.

Nesta revisão serão descritos apenas os modelos exponencial e esférico por terem sido os modelos adotados para a modelagem geoestatística dos dados. Uma discussão mais detalhada sobre a motivação pela adoção destes modelos será apresentada posteriormente neste documento.

Modelo exponencial:

$$\gamma(\|h\|;\theta) = \begin{cases} 0, & se\ \|h\| = 0 \\ \tau^2 + \sigma^2\left[1 - e^{-3\frac{\|h\|}{\phi}}\right], & se\ \|h\| > 0 \end{cases}, \qquad (2.6)$$

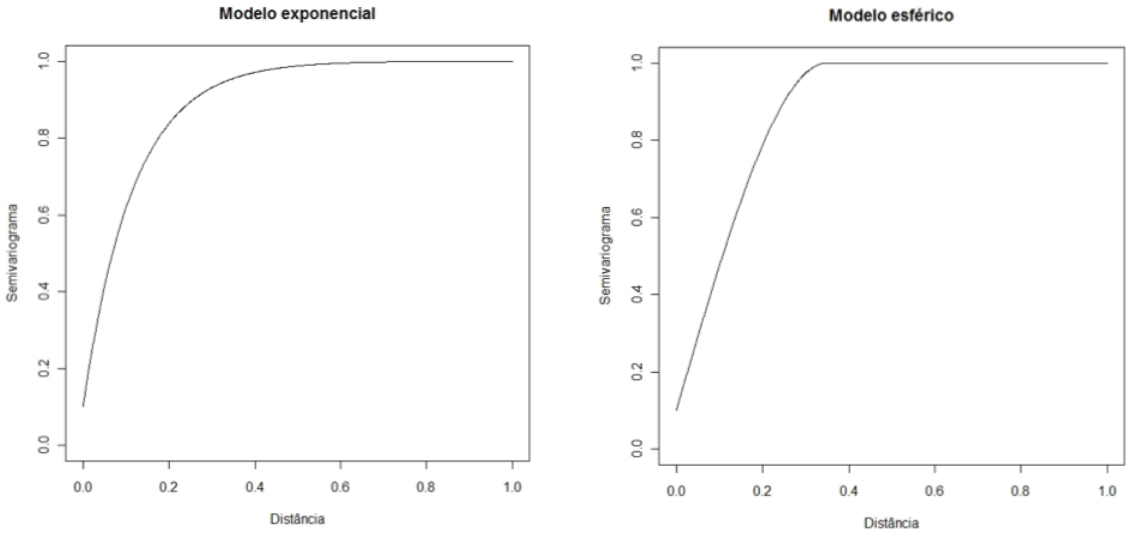
onde $\theta = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Sabendo que, τ^2 é efeito pepita, σ^2 o patamar, e ϕ o alcance.

Modelo esférico:

$$\gamma(\|h\|;\theta) = \begin{cases} 0, & se\ \|h\| = 0 \\ \tau^2 + \sigma^2\left[\frac{3\|h\|}{2\phi} - \frac{\|h\|^3}{2\phi^3}\right], & se\ 0 < \|h\| \leq \phi \\ \tau^2 + \sigma^2, & se\ \|h\| > \phi \end{cases}, \qquad (2.7)$$

onde $\theta = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Sabendo que, τ^2 é efeito pepita, σ^2 o patamar, e ϕ o alcance. Este modelo é válido apenas em $\mathbb{R}$, $\mathbb{R}^2$ e$\mathbb{R}^3$ mas, segundo Soares (2000) é um dos modelos mais utilizados nas aplicações da geoestatística.

Figura 2 - Ilustração das curvas do semivariograma exponencial e esférico.



Fonte: Autor (2012)

2.1.4.5 Métodos de estimação do variograma

Um processo geoestatístico $Z(s)$ pode ter a sua estrutura de dependência descrita através do variograma se for, pelo menos, intrinsicamente estacionário.

Hilário (2009) resume em 2 etapas fundamentais o processo usual de estimação do variograma.

1. obtém-se um conjunto finito de estimativas pontuais da função variograma a partir da amostra do processo, as quais são determinadas em pontos específicos do domínio da função. Alguns autores chamam a este conjunto de estimativas pontuais do variograma, ou variograma empírico.
2. Utilizando o variograma empírico, é para ajustado o modelo (teórico) de variograma mais adequado, e então se estima os parâmetros desse modelo.

Na primeira etapa, objetiva-se encontrar um conjunto finito de estimativas pontuais da função variograma. Matheron (1963) propôs a seguinte expressão como estimador pontual do variograma

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{(s_i, s_j) \in N(h)} (Z(s_i) - Z(s_j))^2 \quad (2.8)$$

onde $N(h)$ é o conjunto de pares de localizações cuja diferença é igual ao vetor $h.$, ou seja,

$$N(h) = \{(s_i, s_j): s_i - s_j = h; i, j = 1, \ldots, n\} \quad (2.9)$$

Cressie (1993) coloca que o estimador clássico proposto por Matheron é não viesado, porém possui propriedades pobres e são afetadas por outliers devido ao termo quadrático. Para contornar este problema Cressie apresenta em seu trabalho uma proposta de estimador mais robusto para a estimação do variograma.

$$2\bar{\gamma}(h) = \frac{\left\{\frac{1}{N(h)} \sum_{(s_i, s_j) \in N(h)} N(h) \, |Z(s_i) - Z(s_j)|^{\frac{1}{2}}\right\}^4}{\left(0.457 + \frac{0.494}{|N(h)|}\right)} \quad (2.10)$$

Posteriormente, inicia-se a segunda etapa, onde é escolhida a família de variogramas, e então são estimados os parâmetros da função do variograma de modo que estes se aproximem o máximo possível das estimativas pontuais do variograma. Os métodos de máxima verossimilhança e de mínimos quadrados se destacam por serem os mais utilizados na estimação. Também, por serem os métodos utilizados nos dados deste trabalho, os mesmos são apresentados a seguir.

Método da Máxima Verossimilhança

Para a estimação por Máxima Verossimilhança, pressupõe-se que o processo Z(s) tem distribuição normal. A grande importância deste método é devida às boas propriedades dos seus estimadores, consistentes e assintoticamente eficientes (DIGGLE, P. J.; RIBEIRO JR., P. J., 2007). Segundo Hilário (2009), se admitido que o processo seja estacionário, e que as observações são provenientes de uma distribuição normal multivariada, com matriz de covariâncias $\Sigma(\theta)$, isto é, que $Z = (Z(s_1), \ldots, Z(s_n)) \sim N(\mu, \Sigma(\theta))$, então o logaritmo da função de verossimilhança é expresso por

$$log\ L(\theta) = -\frac{n}{2}\ log(2\pi) - \frac{1}{2}log\left(det\big(\Sigma(\theta)\big)\right) + \frac{1}{2}(Z-\mu)^T\Sigma(\theta)^{-1}(Z-\mu), \qquad (2.11)$$

onde $det(\Sigma(\theta))$ representa o determinante da matriz $\Sigma(\theta)$. O estimador de máxima verosimilhança, $\hat{\theta}_{ML}$, é obtido determinando o valor $\theta \in \Theta$ que maximiza a função $log\ L(\theta)$.

Método da Máxima Verossimilhança Restrita

O método de máxima verossimilhança restrita quando adotado trata o modelo de forma que cada observação seja dividida em duas componentes independentes, uma referente aos efeitos fixos e outra aos aleatórios. Este método foi desenvolvido por Patterson e Thompson (1974), e corresponde a minimizar a expressão (2.12) para estimar θ

$$log\ L(\theta) = \left(\frac{n-1}{2}\right)\ log(2\pi) + \frac{1}{2}log|A^T\Sigma(\theta)A| + \frac{1}{2}(Z - A^TX\beta)^T(A^T\Sigma(\theta)A)^{-1}(Z - A^TX\beta), \quad (2.12)$$

onde $A = (a_{ij})$ é uma matriz $(n-1)\ x\ n$ aos quais os elementos são:

$$a_{ij} = \begin{cases} 1, para\ i = j\ ; j = 1, \dots, n-1 \\ -1, para\ i = j+1\ ; j = 1, \dots, n-1 \\ 0, caso\ contrário \end{cases}$$

Diggle & Ribeiro Junior (2000) apontam que o estimador de máxima verossimilhança restrita é menos viesado em pequenas amostras, além de serem amplamente recomendados para estimar parâmetros de modelos geoestatísticos.

Método dos Mínimos quadrados

Considere $(2\hat{\gamma}(h_1), \cdots, 2\hat{\gamma}(h_H))$, em que $H \in \mathbb{N}$, um vetor formado pelas estimativas pontuais do variograma, e, $2\gamma(\theta) = (2\gamma(h_1;\theta), \cdots, 2\gamma(h_H;\ \theta))$, outro vetor constituído por componentes definidas de um modelo de variograma paramétrico válido, onde estes vetores foram obtidos dos mesmos pontos $h_1, \cdots, h_H$. O estimador de mínimos quadrados $\hat{\theta}_{LS}$ é determinado pela solução $\theta \in \Theta$ que minimiza uma expressão do tipo (MIRANDA, 2009)

$$\left(2\hat{\gamma} - 2\gamma(\theta)\right)^T \mathbf{V}^{-1}\left(2\hat{\gamma} - 2\gamma(\theta)\right) \quad (2.13)$$

onde $\mathbf{V}$ representa a matriz de covariâncias do estimador.

É denominado estimador de mínimos quadrados simples (denotado por *OLS*) quando $\mathbf{V}$ for a matriz identidade de ordem H; no caso em que $\mathbf{V}$ for uma matriz diagonal tal que $v_i = Var[2\hat{\gamma}(h_i)]$, o método passa a ser designado por mínimos quadrados ponderados (denotado por *WLS*); finalmente, caso $\mathbf{V}$ for uma matriz quadrada tal que $v_{i,j} = Cov\left[2\hat{\gamma}(h_i), 2\hat{\gamma}(h_j)\right]$ o estimador é tratado como estimador de mínimos quadrados generalizados (denotado por *GLS*).

2.1.5 Critérios para Seleção de Modelos

É de grande importância na análise de dados a escolha do modelo apropriado, do ponto de vista estatístico (Bozdangan. H., 1987). Busca-se encontrar um modelo que explique satisfatoriamente o comportamento da variável resposta. Diversos critérios para seleção de modelos são apresentados na literatura. Neste trabalho foram utilizados dois destacados critérios baseados no máximo da função de verossimilhança: o Teste da Razão de Verossimilhança e o Critério de Informação de Akaike (AIC).

Teste da Razão de Verossimilhança

O teste da razão de verossimilhança é adequado no caso de modelos aninhados (quando um modelo é um caso especial do outro). Este método permite comparar dois modelos ajustados por máxima verossimilhança.

Denota-se L_2 a verossimilhança do modelo completo (com a adição de algum(s) fator(es) ou covariável(s)) e L_1 a verossimilhança do modelo restrito. Tem-se que $L_2 > L_1$ e, consequentemente $log\ L_2 > log\ L_1$. A estatística de teste é sempre positiva como apresentado a seguir:

$$2log\left(\frac{L_2}{L_1}\right) = log(L_2) - log(L_1) \quad (2.14)$$

Critério de Informação de Akaike (AIC)

O Critério AIC é definido por:

$$AIC = -2l + 2p \tag{2.15}$$

onde:

l é o ln da função de verossimilhança;

p é o número de parâmetros do modelo considerado.

A decisão da escolha do melhor modelo é realizada avaliando entre os possíveis modelos candidatos o que apresentou menor valor de AIC.

2.1.6 Krigagem

A análise geoestatística não se restringe a modelagem da dependência espacial. Um dos grandes interesses é predizer valores em locais não observados. Esta predição pode ser desde um ou mais pontos específicos, como de uma malha de pontos interpolados que permitam visualizar o comportamento da variável na região. Portanto, para se detalhar a área em estudo faz-se necessário a utilização de um método de interpolação.

Existem diversos métodos de interpolação, tais como: triangulação, método poligonal, médias locais da amostra e inverso do quadrado das distâncias. Porém, apresentam limitações como não considerar anisotropia, estimativas descontínuas entre outras.

A Krigagem é o método de interpolação proposto pela geoestatística. Este interpolador pondera os vizinhos do ponto a ser estimado, obedecendo a critérios de não tendenciosidade e mínima variância (ALMEIDA, 1996). Alguns dos tipos de Krigagem são: simples, ordinária, universal, indicadora e probabilística. Cressie (1993) apresenta maiores detalhes sobre tipos de krigagem. Nos dados deste estudo foi adotado o método de krigagem ordinária.

Sinônimo de predição ótima, a krigagem pode ser descrita como um método de fazer inferência de valores não observados de processo aleatório Z(.) dado por $\{z(s): s \in D \in \Re^d\}$

Na Krigagem ordinária, o conhecimento prévio da média, que é calculado internamente, não se faz necessário. Além disso, apenas a esperança da variável nos pontos desconhecidos é estimada por este método, enquanto os parâmetros de covariância são assumidos como conhecidos.

A predição através do método de Krigagem Ordinária pode ser expressa como uma combinação linear $\widehat{W}(s) = \sum a_i(s)Z_i$. Onde o termo a_i é tratado como predição ponderada, ou krigagem ponderada, e tem a propriedade $\sum a_i(s) = 1$ para qualquer localização.

3 METODOLOGIA

3.1 MATERIAL

Os dados analisados neste trabalho são provenientes de um conjunto de informações disponibilizados ao Núcleo de Assessoria Estatística (NAE/Instituto de Matemática) da UFRGS. Neste estudo foram avaliadas as características da gasometria dos solos, fatores físicos associados, assim como as informações referentes à distribuição geográfica. Para o estudo geoestatístico foram considerados os gases amostrados: Metano, Etano e Eteno.

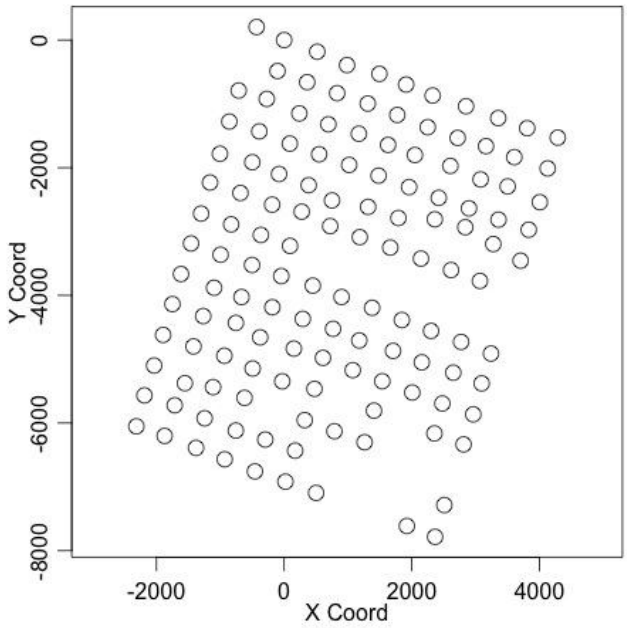
Esta pesquisa não foi acompanhada por um pesquisador especializado na área da geologia que é a pessoa que tem maior conhecimento sobre o fenômeno do comportamento da distribuição espacial dos gases envolvidos, tampouco foi realizado contato com a fonte geradora do conjunto de dados para obtenção de informações quanto à precisão do instrumento de coleta. Logo, neste estudo, tem-se como intuito exclusivo a aplicação das técnicas geoestatísticas ao conjunto de dados. Algumas pré-suposições foram necessárias durante o processo de modelagem e a validação futura das mesmas por um profissional da área poderá contribuir para este estudo.

3.2 LOCALIZAÇÃO GEOGRÁFICA E MÉTODO DE AMOSTRAGEM

Pesquisas na área e exploração e produção de petróleo se caracterizam pelo sigilo de suas informações. Portanto, a localização geográfica do levantamento não será divulgada, e ainda, as coordenadas georeferenciadas tiveram seus pontos rotacionados para garantir maior segurança das informações mediante a publicação deste material.

A amostragem foi realizada, em sua maioria, através da coleta dos dados em uma malha regular, e igualmente espaçada em direções perpendiculares, totalizando 133 unidades amostrais.

Figura 3 - Mapa Amostral



Fonte: Autor (2012)

3.3 ANÁLISE DESCRITIVA

A análise estatística descritiva, como o próprio nome diz, trata de descrever os dados. Considerada uma das três subáreas da estatística, juntamente com a probabilística e a inferencial, a análise descritiva tem como objetivo resumir a informação constante nos dados permitindo uma visão global do comportamento e variação desses valores, além de organizar e descrever por meio de tabelas, gráficos e medidas descritivas.

As medidas descritivas são classificadas como: medidas de posição, medidas de dispersão, medidas de assimetria e de curtose.

Os resultados obtidos através da análise descritiva permitem também a tomada de decisão quanto ao uso de transformações na variável resposta. Busca-se, sempre que possível, obter através das transformações distribuições com boas características de forma e simetria, além de um tratamento quanto a valores atípicos (*outliers*). Diggle & Ribeiro (2007) afirmam que a classe de modelos gaussianos pode ser estendida através de uma transformação da variável resposta original, e com isso uma maior flexibilidade no modelo, onde este geralmente fornece um bom ajuste aos dados empíricos.

3.4 ANÁLISE EXPLORATÓRIA ESPACIAL

A característica fundamental da análise exploratória espacial é estudar os valores observados levando em consideração sua localização geográfica de forma a avaliar o comportamento espacial de uma variável sobre cada eixo de coordenadas, podendo assim auxiliar na identificação da existência de alguma tendência na dependência espacial da mesma.

3.5 ANÁLISE GEOESTATÍSTICA

O processo de análise geoestatística dos gases em estudo inicia-se pela investigação, através do método empírico, da dependência espacial, além de justificar a adoção de uma estrutura de covariância, por meio da que melhor se adapta aos resultados obtidos. Nesta etapa são avaliadas diversas configurações do variograma empírico, testando diferentes configurações de *bins* e distância máxima. Encontrado o variograma *bin* que justifique (graficamente) a adoção de uma estrutura de correlação espacial, parte-se para etapa de ajuste, onde os resultados obtidos são utilizados simplesmente como argumentos para estimação através dos métodos de estimação por máxima verossimilhança.

Em seguida são realizados os ajustes através dos métodos de máxima verossimilhança e máxima verossimilhança restrita utilizando as estruturas de covariância adotadas na etapa de modelagem empírica. Nesta fase, para cada combinação de método de estimação e estrutura de covariância, são realizados os ajustes para o modelo constante (sem covariável) e também com cada covariável separadamente. Posteriormente são adotados os critérios para seleção de cada modelo.

3.6 CRITÉRIO PARA SELEÇÃO DE MODELOS

Como critério para a seleção de modelos é utilizado, primordialmente, o teste da razão de verossimilhanças, visando comparar um modelo com a inclusão de uma covariável, que no caso são os fatores de umidade, tipo de solo, cor e uso do solo, com o modelo sem covariável, conforme descrito no item 2.1.5.

O critério de informação de Akaike é também utilizado com o objetivo de confrontar o resultado alcançado pelo teste da razão de verossimilhança.

3.7 PREDIÇÕES

Uma vez estimados os parâmetros para os modelos, com suas respectivas estruturas de covariância através dos métodos máxima verossimilhança e máxima verossimilhança restrita, é dado o momento realizar inferências sobre a superfície em locais não observados. A krigagem ordinária foi o método de interpolação empregado para obter os resultados das predições e variância das predições.

3.8 VALIDAÇÃO CRUZADA

Para verificar a adequabilidade da modelagem da distribuição dos gases metano, etano e eteno, é utilizado o método de validação cruzada a fim de verificar a fidedignidade das informações obtidas da amostra.

3.9 SOFTWARES UTILIZADOS

Para a realização das análises gráficas e estatísticas para os métodos propostos foi utilizado o software R na versão 2.15 for MAC OS em conjunto com o pacote/biblioteca *geoR* versão 1.7.4 (Diggle & Ribeiro Junior, 2000,2001) disponibilizado gratuitamente através do website *www.r-project.org*.

4 ANÁLISE GEOESTATÍSTICA

A estratégia adotada neste trabalho para realização da análise geoestatística das variáveis em estudo: metano, etano e eteno; foi de analisar separadamente cada um dos gases. Portanto, são apresentados neste capítulo três subcapítulos dedicados à análise de cada gás. Tendo isso em mente, é adotada a metodologia de análise já descrita para conduzir o estudo dos gases metano, etano e eteno.

Inicialmente, são apresentadas breves informações descritivas referentes ao conjunto de dados amostrado, tais como resumo espacial das coordenadas X e Y.

As informações de máximo, mínimo e da amplitude referentes às coordenadas de X e Y são apresentadas na Tabela 1. Posteriormente, a Tabela 2 apresenta o tamanho da amostra, que é de 133 unidades amostrais, juntamente a menor e maior distância entre os pares.

Tabela 1 - Informação das coordenadas.

	Coordenadas	
	CoordX	CoordY
Mínimo	-2315	-7783
Máximo	4286	204
Amplitude	6601	7987

Fonte: Autor (2012)

Tabela 2 - Tamanho da amostra e distância entre pares de observações.

Amostra	
Número de observações	133
Menor distância entre pares	301,63
Maior distância entre pares	8462,26

Fonte: Autor (2012)

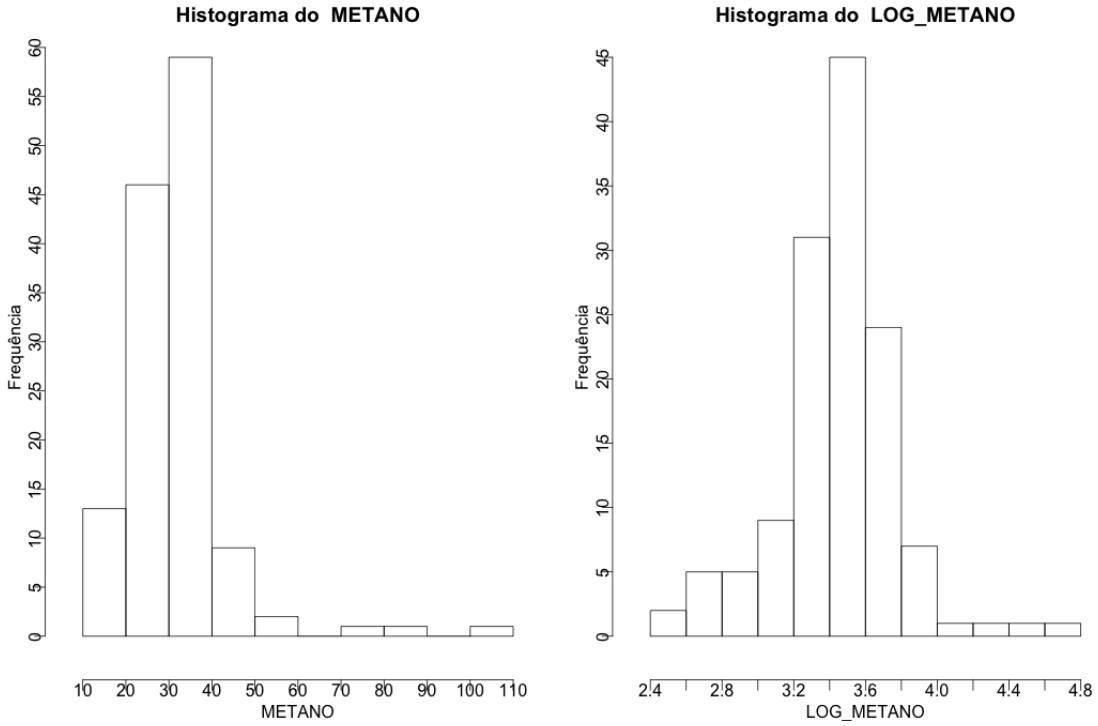
4.1 ANÁLISE DA VARIÁVEL METANO

4.1.1 Análise Descritiva

A estatística descritiva é uma técnica exploratória que objetiva resumir a informação contida nos dados de maneira a organizar e descrever através de gráficos, tabelas, medidas de posição e dispersão, contribuindo com uma visão geral da sua variação.

A assimetria para os dados do METANO fica visível através do histograma da Figura 4 (esquerda). Na tentativa de diminuir esta assimetria foi realizada a transformação $ln(x+1)$ nos dados originais do metano criando-se uma nova variável denominada LOG_METANO. Após a transformação, pode-se visualizar claramente através do histograma do LOG_METANO (Figura 4, à direita) uma distribuição bem mais comportada quanto à assimetria.

Figura 4 - Representação da distribuição dos dados originais do METANO (à esquerda) e da distribuição dos dados transformados LOG_METANO (à direita).

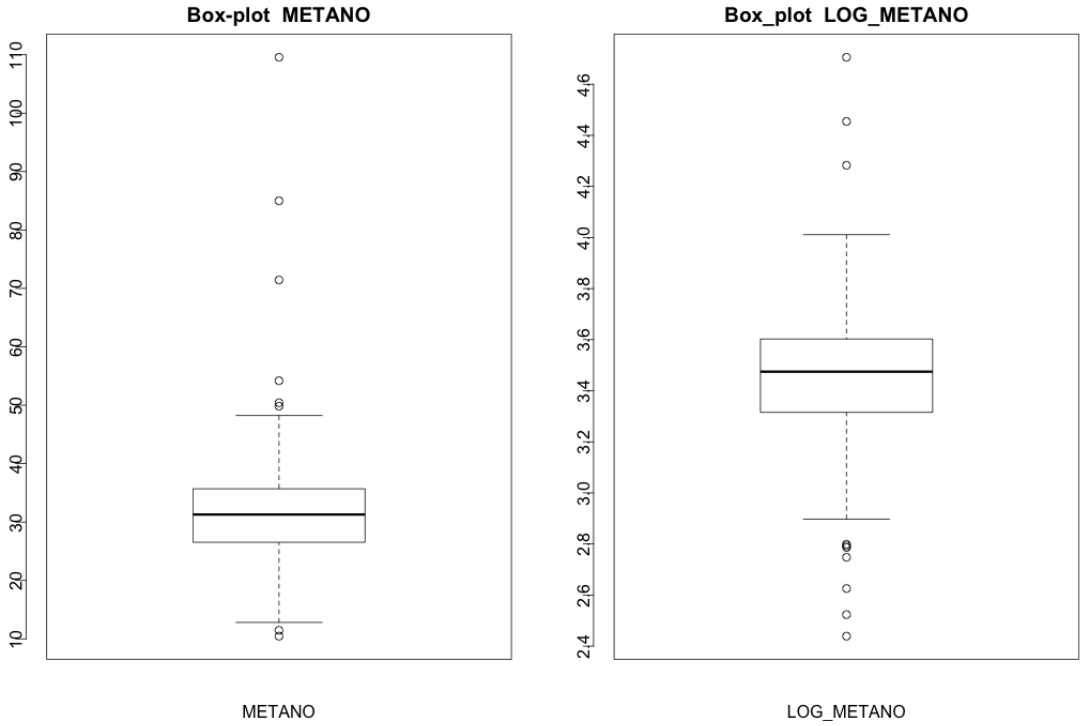


Fonte: Autor (2012)

A Figura 5 apresenta o Box-plot para às variáveis: METANO e LOG_METANO, respectivamente. O Box-plot é um gráfico que também colabora, além de ser mais informativo, na análise da posição, dispersão e identificação de valores extremos. Este gráfico é construído usando a informação dos quartis formando um retângulo, onde a parte inferior e superior deste são sempre o percentil 25 e 75 (os quartis superiores e inferiores, respectivamente). A linha ao meio da caixa representa o percentil 50 (mediana). As linhas que saem do retângulo percorrem até os valores mais atípicos. Para os dados do METANO, o gráfico mostra valores mais afastados da mediana do que os dados relativos ao

LOG_METANO. Nos dados transformados a dispersão apresenta uma leve diminuição, apesar de valores atípicos serem claramente observados.

Figura 5 - Box-plot para os dados da variável METANO (à esquerda) e para os dados da variável LOG_METANO (à direita).



Fonte: Autor (2012)

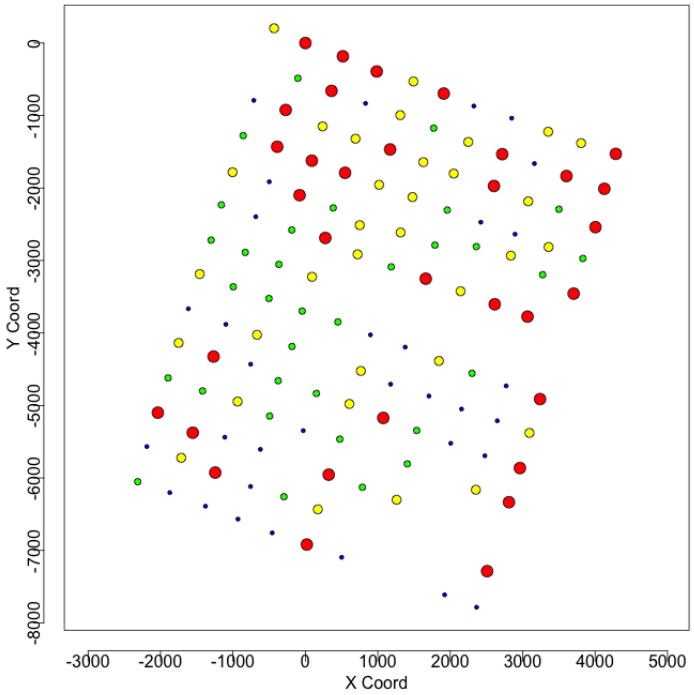
Devido a variável LOG_METANO apresentar melhores características em sua distribuição, será dado continuidade às análises somente sobre os dados transformados.

4.1.2 Análise Exploratória Espacial

A análise descritiva fornece de uma maneira muito básica informações sobre o comportamento da dispersão dos dados. Além disso, não considera a distribuição dos dados no espaço. Explorar os valores observados levando em consideração sua localização geográfica é o primeiro passo da análise exploratória espacial. Uma das formas de visualizar graficamente a distribuição dos dados é através da associação de cores fortes aos maiores valores da variável. A função *points()* do pacote *geoR* permite a análise visual da concentração do LOG_METANO referente a cada localização geográfica amostrada, onde as cores e tamanho dos pontos (círculos)

plotados no gráfico representam cada quartil do conjunto de dados. Círculos menores e de cor azul correspondem ao primeiro quartil, seguindo pela cor verde, amarelo e vermelho que se referem ao segundo, terceiro e quarto quartis, respectivamente. Na Figura 6 é apresentado o gráfico da localização geográfica da variável LOG_METANO.

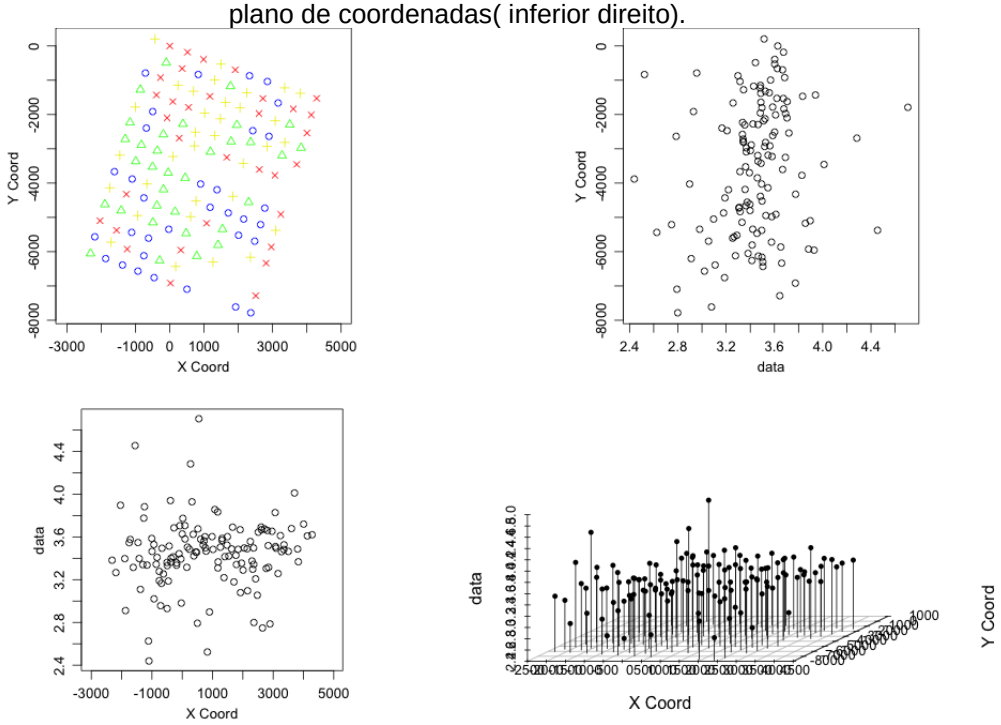
Figura 6 - Localização geográfica do LOG_METANO associando cores fortes com a magnitude dos dados.



Fonte: Autor (2012)

Avaliar o comportamento espacial da variável sobre cada eixo de coordenadas pode auxiliar na identificação da existência de alguma tendência na dependência espacial da mesma. Para ambas as coordenadas, pode-se observar graficamente (Figura 7, superior direito para coordenada Y e inferior esquerdo para a coordenada X) que as informações não evidenciam a presença de tendência.

Figura 7 - Localização geográfica do LOG_METANO (superior esquerdo), valores do LOG_METANO versus as coordenadas (superior direito e inferior esquerdo), e valores do LOG_METANO sobre o plano de coordenadas(inferior direito).



Fonte: Autor (2012)

4.1.3 Análise Geoestatística

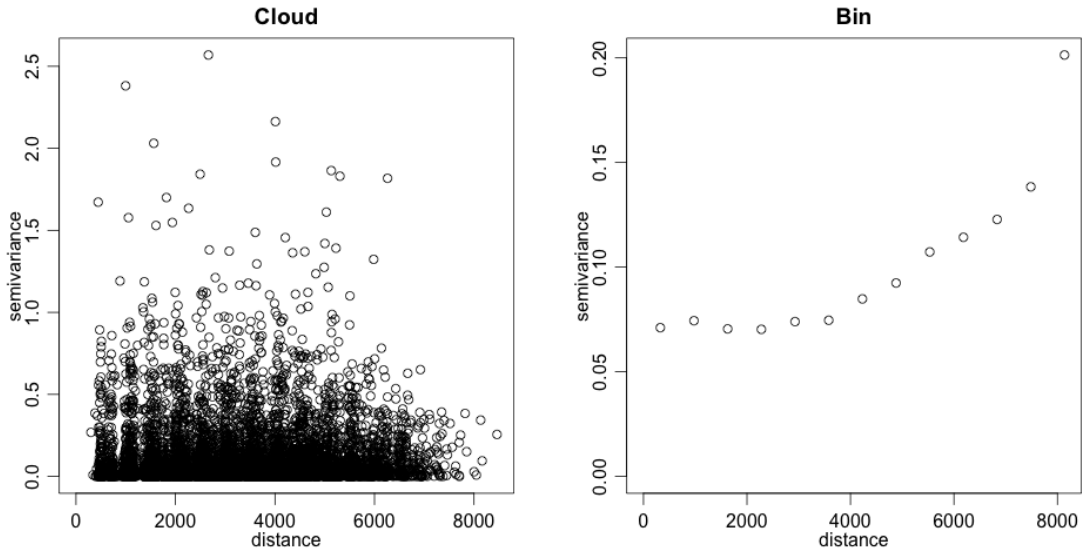
Neste trabalho, ao estudar a dependência espacial para os dados do LOG_METANO, foram utilizados métodos exploratórios para auxiliar na identificação de características que justifiquem a escolha de um modelo que represente a estrutura de covariância espacial. É de interesse verificar a influência dos fatores físicos na caracterização da distribuição espacial do LOG_METANO. Para isso, foi investigado a associação entre os dados e as covariáveis: Umidade, Tipo de Solo, Cor e Uso do Solo, adicionando cada uma das mesmas, como um fator de tendência na modelagem geoestatística.

O início da análise geoestatística deu-se através do uso de métodos exploratórios tais como os variogramas empíricos de nuvem (*cloud*) e pontuais (*bins*) (Figura 8). O variograma empírico pode ser utilizado como uma ferramenta que visa identificar a correlação das observações à medida que a distância aumenta entre elas. Assumindo que o processo gerador é gaussiano estacionário, a correlação entre os pontos deve depender somente da distância entre eles, e, com o aumento desta distância, a correlação tende a zero. Além de sua utilização como ferramenta

exploratória, o variograma também é utilizado no ajuste da função de covariância paramétrica, dado que é um estimador não viesado do variograma teórico. Porém, devido a cada ponto entrar diversas vezes no cálculo das distâncias em função dos diferentes números de pares que geram cada ponto, os variogramas empíricos são muito erráticos em certos cenários, tais como na existência de valores atípicos, ou ainda, no caso de a malha amostral não ser capaz caracterizar toda a dependência espacial do fenômeno.

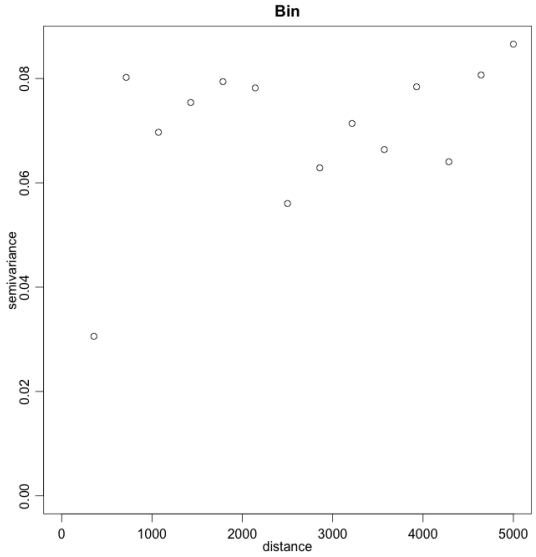
O gráfico de dispersão de todas as distâncias entre dois pontos observados em qualquer direção (omnidirecional) contra o seu referente valor do semivariograma é apresentado na Figura 8 (variograma *cloud*). Nele, podem-se visualizar características que auxiliam no processo de modelagem, tais como a amplitude e a variância do processo observado. Quanto ao variograma *bins*, que corresponde à média calculada para cada *lag* do variograma *cloud*, foi utilizado o estimador robusto (*modulus*) omnidirecionalmente, e pode-se verificar visualmente que o mesmo não apresenta um formato com características que se assemelhem às dos modelos usuais. Porém, observa-se certa estabilidade no comportamento da semivariância até aproximadamente à distância 4000. Como esta análise foi realizada sobre toda a amplitude das distâncias observadas, uma análise mais detalhada até a distância 5000 é apresenta na Figura 9.

Figura 8 – Variograma Cloud Omnidirecional(esquerda) e o variograma Bin Omnidirecional (direita) da variável LOG_METANO.



Fonte: Autor (2012)

Figura 9 - Variograma Bin com distância máxima de 5000.



Fonte: Autor (2012)

Devido a sua forma, o gráfico apresentado na Figura 9 justifica a escolha dos modelos exponencial e esférico para representar a estrutura de covariância ajustada aos pontos do variograma dos dados da variável LOG_METANO. Os métodos de estimação por mínimos quadrados ordinários (OLS) e mínimos quadrados ponderados (WLS) foram utilizados inicialmente para ajustar cada um dos modelos. A função *variofit*() foi utilizada para estimação dos parâmetros. No processo de modelagem, foi necessário fixar o valor do efeito pepita (*nugget*) para que a função

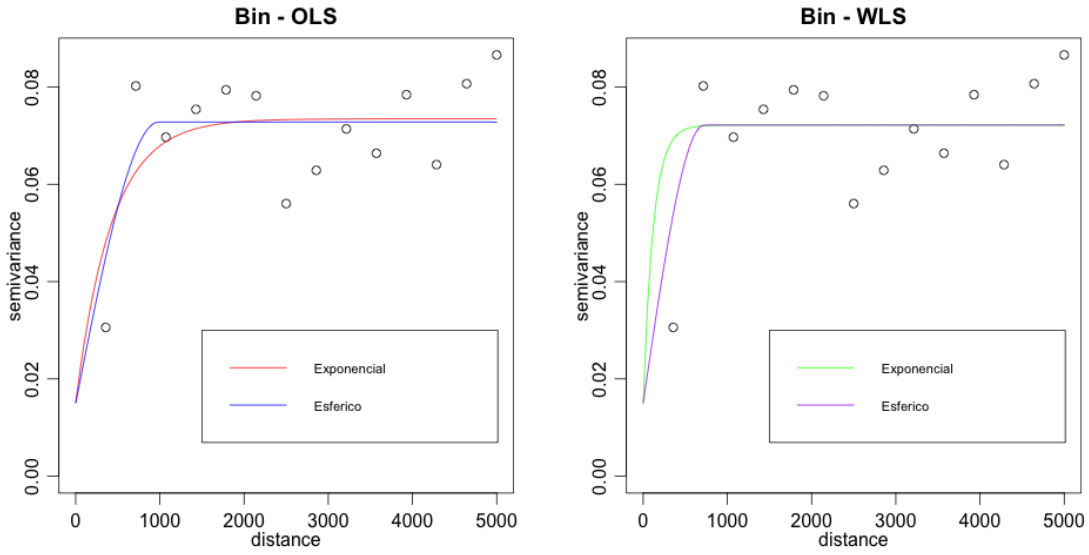
de estimação *variofit()* obtivesse estacionariedade no processo, além disso foram especificados os valores iniciais idênticos para os quatro modelos investigados. Os resultados obtidos na estimação dos parâmetros são apresentados na Tabela 3 e os ajustes obtidos estão plotados na Figura 10.

Tabela 3 - Resultado das estimativas dos parâmetros para os dados do LOG_METANO.

Estimadores	OLS		WLS	
Parâmetros	Exponencial	Esférico	Exponencial	Esférico
Nugget	0,0150	0,0150	0,0150	0,0150
Sill	0,0585	0,0578	0,0572	0,0572
Range	425,2340	983,7574	118,8874	722,4949

Fonte: Autor (2012)

Figura 10 - Variogramas ajustados utilizando os métodos de mínimos quadrados ordinários (OLS) e mínimos quadrados ponderados (WLS) pelos modelos exponencial e esférico.



Fonte: Autor (2012)

A escolha de um método que proporcione maior precisão na estimação dos parâmetros da função de covariância que representa a estrutura espacial é essencial para a realização de uma predição confiável. Os métodos paramétricos são mais confiáveis quanto à qualidade de estimação, principalmente os métodos que utilizam os estimadores de máxima verossimilhança (DIGGLE e RIBEIRO JÚNIOR, 2000), pois se ajustam a todos os valores amostrados (variograma *cloud*). Por este motivo foi buscado a obtenção de modelos estimados através destes métodos.

Nesta etapa da modelagem, os métodos de máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) foram utilizados para estimação dos

parâmetros do variograma. Deve-se destacar que estes dois métodos utilizam toda a “nuvem” e que as estimativas pontuais do variograma não consideradas de forma isolada na maximização da função de verossimilhança. Na prática, para os métodos de máxima verossimilhança, as estimativas pontuais contribuem apenas para selecionar o modelo de variograma mais adequado.

Neste estudo, tem-se como objetivo testar a associação das covariáveis Umidade, Tipo de Solo, Cor e Uso do Solo na distribuição espacial do LOG_METANO. Ao adicionar estas covariáveis ao vetor de parâmetros do modelo, os estimadores de máxima verossimilhança e máxima verossimilhança restrita possibilitam testar a significância de sua contribuição ao modelo.

Os modelos exponencial e esférico, por terem sido os que melhor se ajustaram às estimativas pontuais, serão os utilizados na modelagem através dos métodos ML e REML. Foram então realizados ajustes ao modelo sem covariável e com cada covariável separadamente. Por fim, os resultados das estimativas de cada ajuste são apresentados e, para os modelos selecionados, são realizadas as validações cruzadas.

A função *likfit()* do pacote *geoR* foi utilizada para estimação dos parâmetros, tanto para o método ML como REML. Os valores iniciais para os parâmetros patamar parcial (*partial sill*) e alcance (*range*) são solicitados pela função (argumento *ini.cov.pars*), assim como o valor do efeito pepita (nugget), onde este último é opcional e pode ser fixado (argumento *fix.nugget*). Os resultados obtidos através da estimação por mínimos quadrados realizada anteriormente auxiliaram na escolha destes valores. Quando utilizado somente estes argumentos nas tentativas de modelagem, a função não conseguiu gerar resultados aceitáveis, onde em certas situações retornava a estimativa para *phi* com o valor zero, e em outras o resultado da estimativa de *phi* excedia por mais de 10 vezes o tamanho da malha amostral. Como alternativa, a função *likfit()* oferece a possibilidade de definir limites de estimação para os valores dos parâmetros *phi (range)* e/ou *sigmasq (partial sill)* através do uso do argumento *limits,* auxiliado pela função *pars.limits()*.Para os dados da variável LOG_METANO, foi optado utilizar o argumento *limits()* delimitando os valores de *sigmasq*.

Para a variável LOG_METANO, foram obtidos os valores para as estimativas dos parâmetros, AIC e logaritmo da função de verossimilhanças, através da utilização do estimador de máxima verossimilhança com estrutura de covariância

exponencial, para os modelos sem covariável(constante) e com cada uma das covariáveis: Umidade, Tipo de Solo, Cor e Uso do Solo (Tabela 4).

Na Tabela 4 pode-se verificar que os ajustes aos modelos 3 e 5 não conseguiram estimar *sigmasq* e *phi*, logo os mesmos foram descartados da análise. Para os demais modelos, o ajuste apresentou estimativas para *phi* iguais aos parâmetros iniciais repassados à função *likfit()*, porém, como colocados anteriormente, ao delimitar os possíveis valores de *sigmasq*, o ajuste conseguiu gerar resultados aceitáveis. Os resultados do teste da razão de verossimilhanças, quando comparado os modelos com a covariável e o modelo sem a covariável (constante), foram menores que o valor crítico 5,99 ($\chi^2_{(2;0,05)}$), portanto, ao nível de significância de 5%, conclui-se que os modelos com covariável não foram significativamente melhores que o modelo 1. O critério de informação de Akaike (AIC) também confirma o modelo 1 como melhor ajuste com o menor valor AIC entre os três ajustes dados como válidos.

Tabela 4 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_METANO, com estrutura de covariância exponencial, utilizando os estimadores de máxima verossimilhança.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Método	ML	ML	ML	ML	ML
Modelo ajuste	Exponencial	Exponencial	Exponencial	Exponencial	Exponencial
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,015	0,015	0,0999	0,015	0,098
PartialSill	0,1949	0,1931	0	0,1915	0
Phi	999,9976	999,9977	0	999,9979	0
AIC	108,2187	111,1178	80,5662	112,4216	81,9674
LogL	-51,1094	-50,5589	-35,2831	-50,2108	-33,9837
Razão de verossimilhança		1,101	31,6526	1,7972	34,2514
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

Utilizando o estimador de máxima verossimilhança restrita com estrutura de covariância exponencial, novamente o método apresentou dificuldades para gerar as estimativas para *phi*, estimando em todos os cinco modelos ajustados o valor

passado como parâmetro inicial solicitado pela função *likfit()*. O modelo 8 foi o que apresentou menor valor estimado para *sigmasq*. Avaliando os testes da razão de verossimilhanças, o modelo 8 apresentou valor maior que 5,99, que significa que, ao nível de significância de 5%, o ajuste deste modelo foi significativamente melhor do que o modelo 6 que se refere ao modelo sem covariável. O modelo 8 também é o que apresenta menor valor AIC, confirmando-o como melhor ajuste.

Tabela 5 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_METANO, com estrutura de covariância exponencial, utilizando os estimadores de máxima verossimilhança restrita.

	Modelo 6	Modelo 7	Modelo 8	Modelo 9	Modelo 10
Método	REML	REML	REML	REML	REML
Modelo ajuste	Exponencial	Exponencial	Exponencial	Exponencial	Exponencial
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,015	0,015	0,015	0,015	0,015
PartialSill	0,1971	0,1993	0,1868	0,1997	0,1931
Phi	999,9976	999,9974	999,9985	999,9974	999,998
AIC	105,5369	108,5042	101,1053	109,5609	108,183
LogL	-49,7684	-49,2521	-45,5526	-48,7805	-47,0915
Razão de verossimilhança		1,0326	8,4316	1,9758	5,3538
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

Os resultados dos ajustes utilizando estimadores de máxima verossimilhança com estrutura de covariância esférica são apresentados na Tabela 6. Pode-se verificar que o método gerou estimativas para *phi* menores que a menor distância entre pares, exceto para o modelo 15. Este resultado implica na não existência de dependência espacial no processo, fato que foi visualmente desconsiderado quando observada a Figura 6. Portanto, será dado seguimento às análises utilizando o ajuste realizado no modelo 15.

Tabela 6 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_METANO, com estrutura de covariância esférica, utilizando os estimadores de máxima verossimilhança.

	Modelo 11	Modelo 12	Modelo 13	Modelo 14	Modelo 15
Método	ML	ML	ML	ML	ML
Modelo ajuste	Esférico	Esférico	Esférico	Esférico	Esférico
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,015	0,015	0,015	0,015	0,015
PartialSill	0,0874	0,0858	0,0849	0,084	0,083
Phi	79,4918	301,6278	193,2761	299,4311	447,4025
AIC	79,825	81,6855	80,5662	81,2804	82,0431
LogL	-36,9125	-35,8428	-35,2831	-34,6402	-34,0215
Razão de verossimilhança		2,1394	3,2588	4,5446	5,782
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

Na Tabela 7 são apresentados os resultados dos ajustes utilizando estimadores de máxima verossimilhança restrita com estrutura de covariância esférica. Com exceção do modelo 18, os demais ajustes apresentaram estimativas para *phi* menores que a menor distância entre pares indicando a não existência de dependência espacial no processo, fato que esta sendo visualmente desconsiderado quando observada a Figura 6. Portanto, se dará seguimento às análises utilizando o ajuste realizado no modelo 18.

Tabela 7 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_METANO, com estrutura de covariância esférica, utilizando os estimadores de máxima verossimilhança restrita.

	Modelo 16	Modelo 17	Modelo 18	Modelo 19	Modelo 20
Método	REML	REML	REML	REML	REML
Modelo ajuste	Esférico	Esférico	Esférico	Esférico	Esférico
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,015	0,015	0,015	0,015	0,015
PartialSill	0,0882	0,0881	0,1107	0,0871	0,0868
Phi	248,3492	140,6489	999,9986	207,2399	257,0306
AIC	80,262	83,022	96,3534	83,1198	84,297
LogL	-37,131	-36,511	-43,1767	-35,5599	-35,1485
Razão de verossimilhança		1,24	-12,0914	3,1422	3,965
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

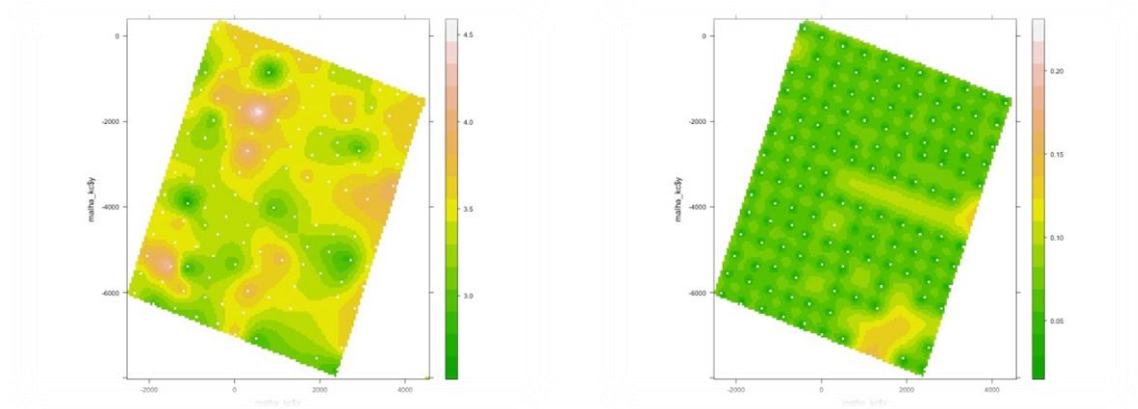
4.1.4 Predições

Para cada um dos métodos de estimação de máxima verossimilhança, ML e REML, quando ajustados por modelos exponencial e esférico, um dos modelos ajustados foi selecionado. Escolhidos estes modelos, que irão representar a estrutura de dependência espacial da variável LOG_METANO, é dado o momento de realizar as predições para os locais não amostrados. A krigagem ordinária foi o método empregado para obter os resultados das predições e variância das predições do LOG_METANO.

Os resultados das predições e variância das predições dos modelos para a variável LOG_METANO, para cada um dos quatro modelos selecionados, estão apresentados abaixo: (Figura 11, Figura 12, Figura 13 e Figura 14)

Figura 11 - Resultados da predição (a) e variância da predição (b) para o LOG_METANO, referentes ao ajuste do modelo 1.

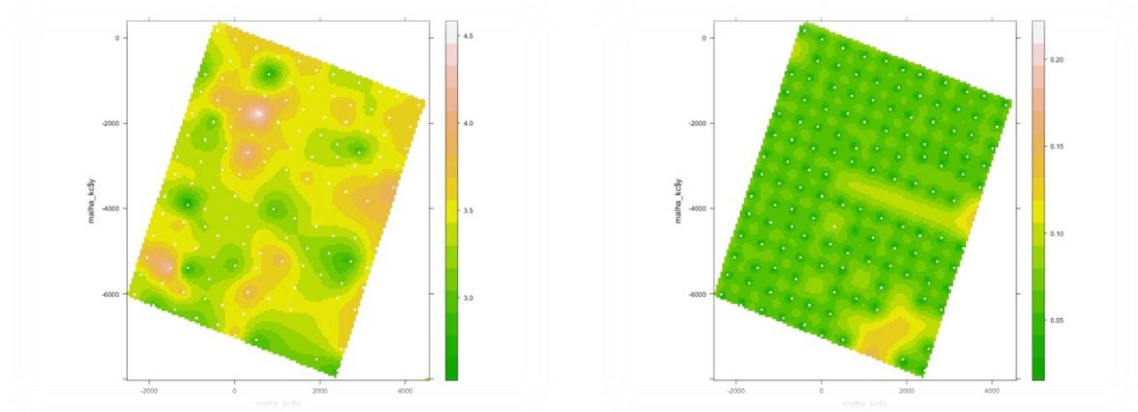
a) b)



Fonte: Autor (2012)

Figura 12 - Resultados da predição (a) e variância da predição (b) para o LOG_METANO, referentes ao ajuste do modelo 8.

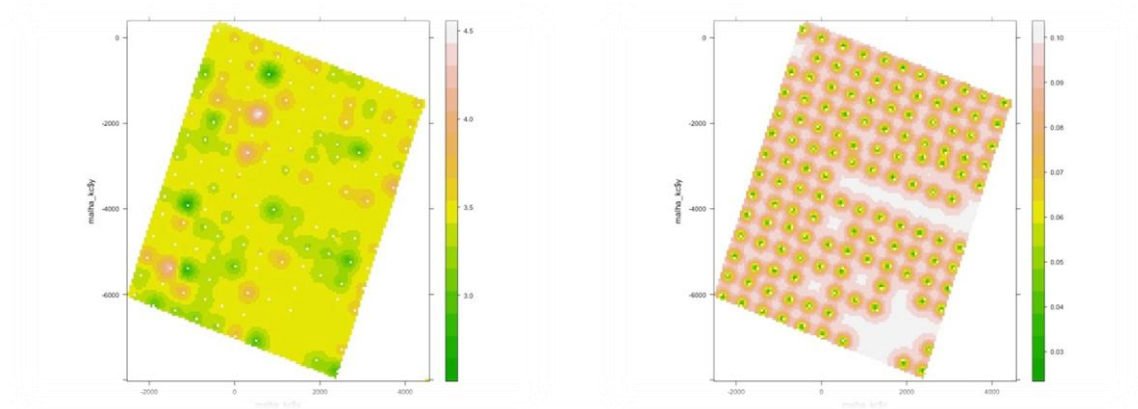
a) b)



Fonte: Autor (2012)

Figura 13 - Resultados da predição (a) e variância da predição (b) para o LOG_METANO, referentes ao ajuste do modelo 15.

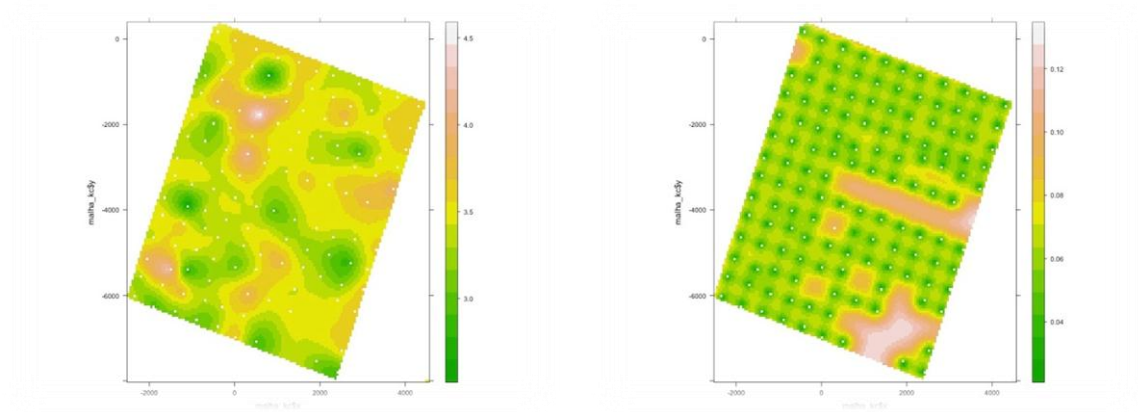
a) b)



Fonte: Autor (2012)

Figura 14 - Resultados da predição (a) e variância da predição (b) para o LOG_METANO, referentes ao ajuste do modelo 18.

a) b)



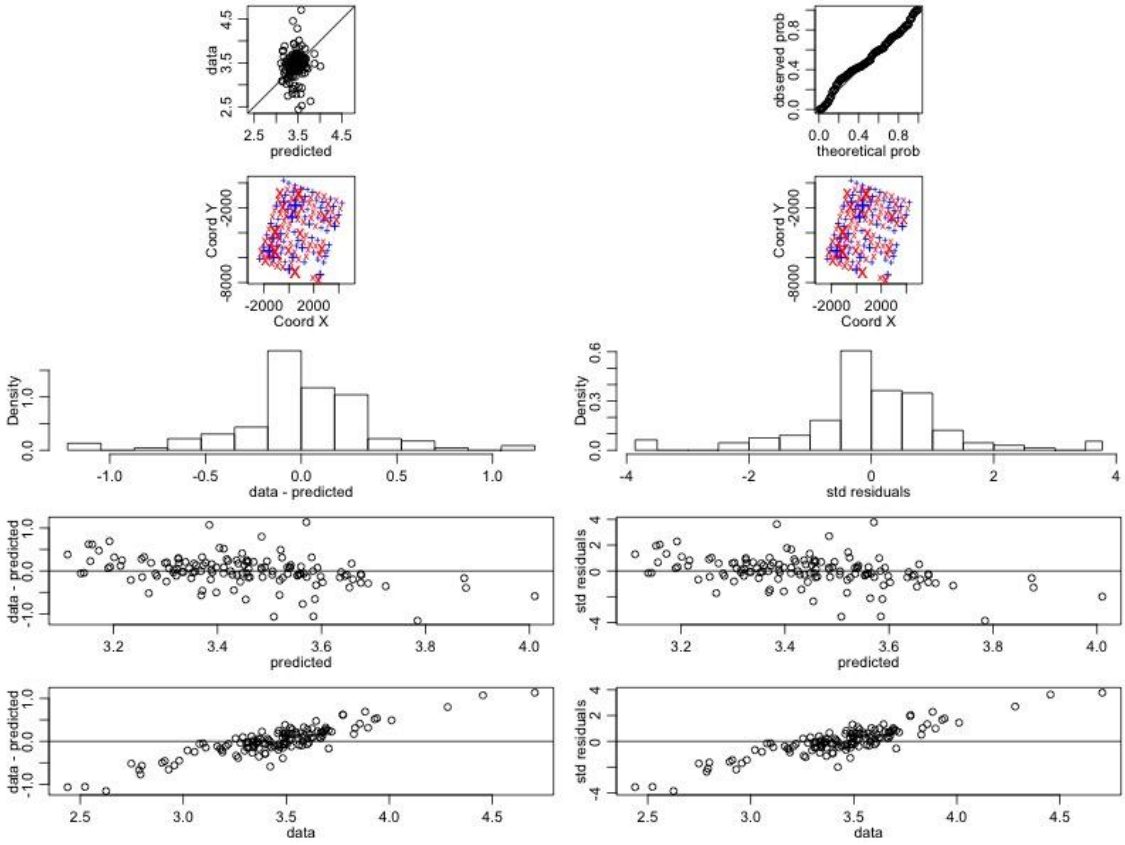
Fonte: Autor (2012)

O estimador de máxima verossimilhança restrita é consagrado na literatura pela qualidade de suas estimativas por serem não viesadas. O modelo 18, estimado por REML com estrutura de covariância esférica, apresentou predições bem definidas e variâncias associadas às predições menores quando comparadas aos demais modelos. Entre os modelos testados, acredita-se que este modelo é o que melhor representa a estrutura espacial dos dados do LOG_METANO.

4.1.5 Validação cruzada

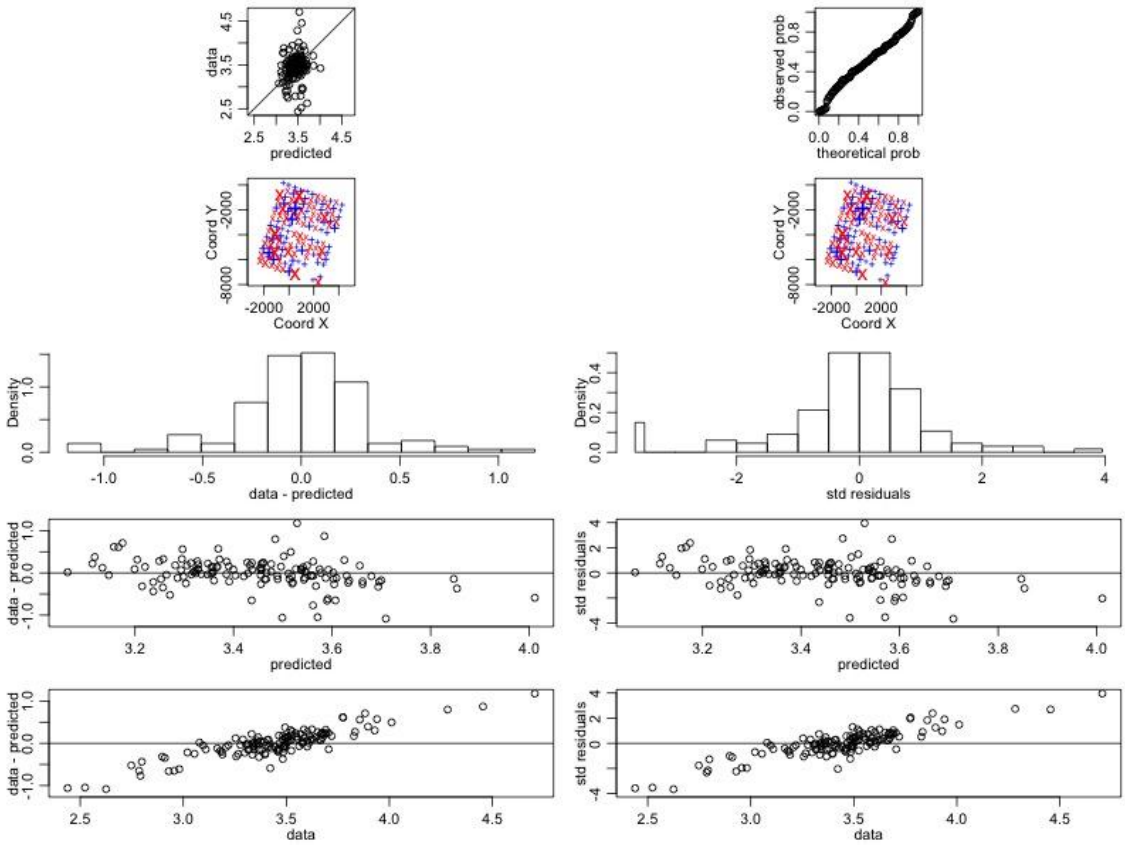
A validação cruzada para os modelos selecionados, em geral, indicaram que os dados do LOG_METANO seguiram uma distribuição normal padronizada. Verificou-se, ainda, que a distribuição dos erros positivos e negativos se apresenta dispersa pela região de estudo, com exceção do modelo 15 onde é constatado um padrão na dispersão dos pontos. Os dados da probabilidade teórica e observada encontram-se sobre a reta, indicando que a krigagem ordinária foi eficiente para a predição. Os *outliers* dificultam a interpretação do gráfico dos valores preditos sobre a reta padrão, pois ao aplicar a validação cruzada, a remoção de uma localização observada onde o valor amostrado representa um *outlier*, o valor predito para este ponto é trazido para próximo do valor observado em seus vizinhos e, consequentemente, gerando uma predição mais próxima à média.

Figura 15 - Validação cruzada para o LOG_METANO referente ao modelo 1.



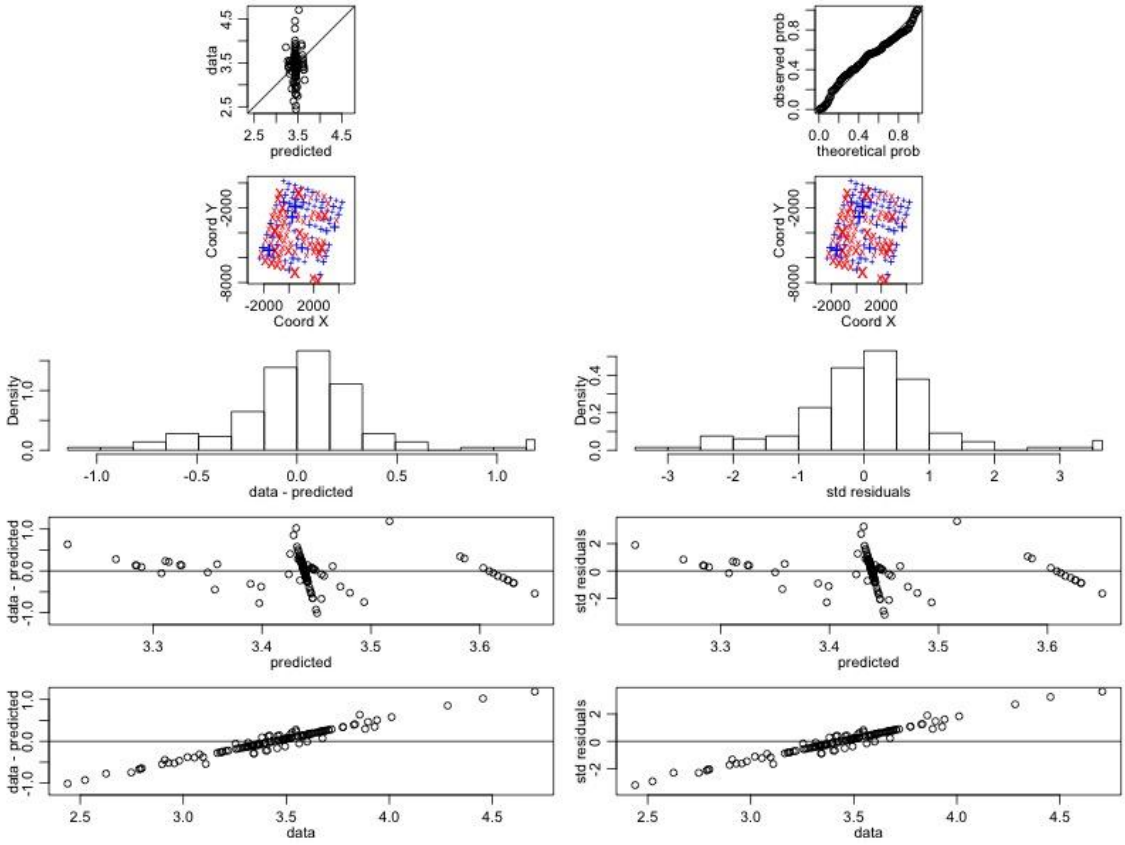
Fonte: Autor (2012)

Figura 16 - Validação cruzada para o LOG_METANO referente ao modelo 8.



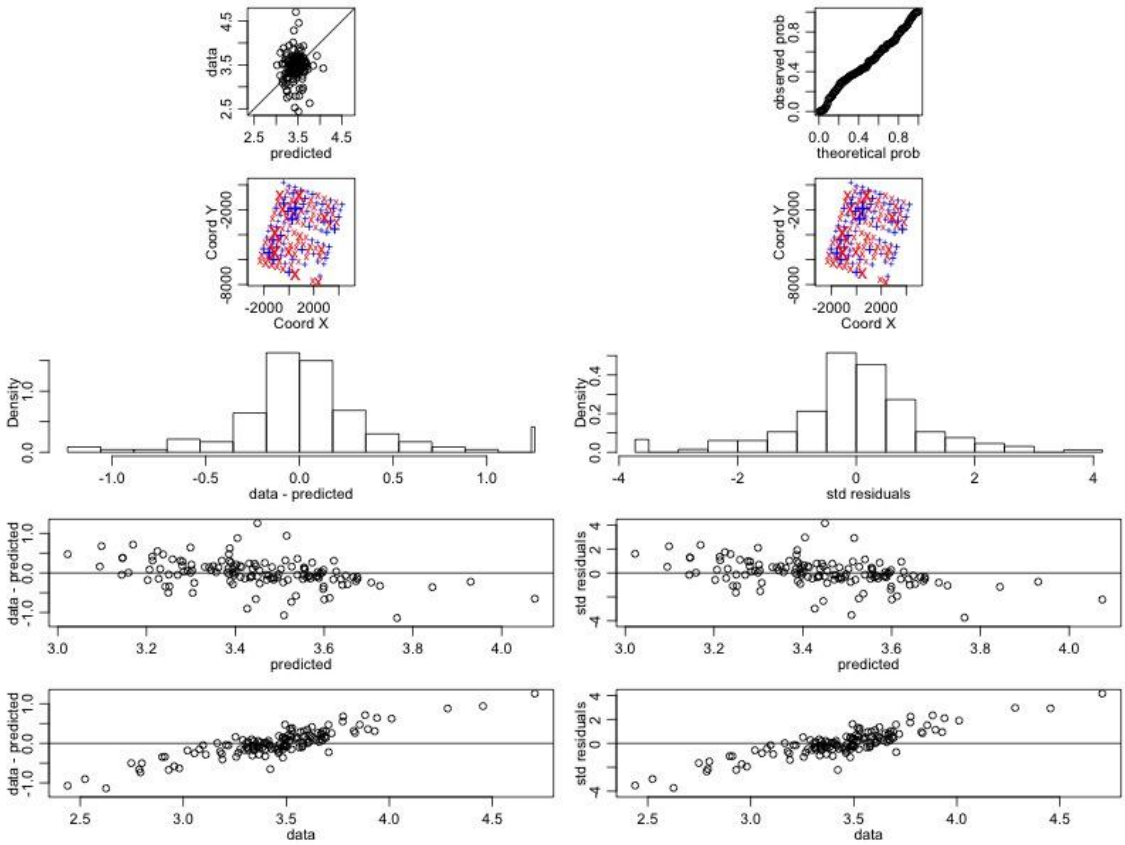
Fonte: Autor (2012)

Figura 17 - Validação cruzada para o LOG_METANO referente ao modelo 15.



Fonte: Autor (2012)

Figura 18 - Validação cruzada para o LOG_METANO referente ao modelo 18.



Fonte: Autor (2012)

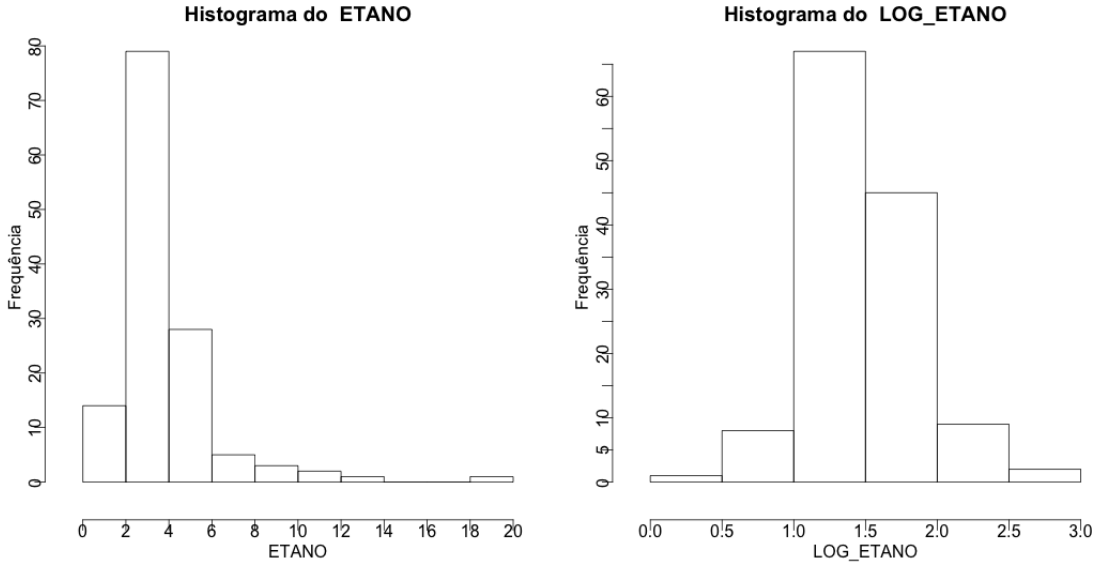
4.2 ANÁLISE DA VARIÁVEL ETANO

Os procedimentos adotados como metodologia de análise para variável ETANO são os mesmos anteriormente utilizados.

4.2.1 Análise Descritiva

Inicialmente, verifica-se que os dados referentes à variável ETANO também apresentam assimetria em sua distribuição como verificado na Figura 19 (esquerda). A transformação $ln(x+1)$ é também aplicada aos dados originais do ETANO, e assim denomina-se a nova variável como LOG_ETANO. Após a transformação, é visualmente perceptível a melhora na assimetria, conforme histograma LOG_ETANO (Figura 19 à direita).

Figura 19 - Representação da distribuição dos dados originais do ETANO (à esquerda) e da distribuição dos dados transformados LOG_ETANO (à direita).

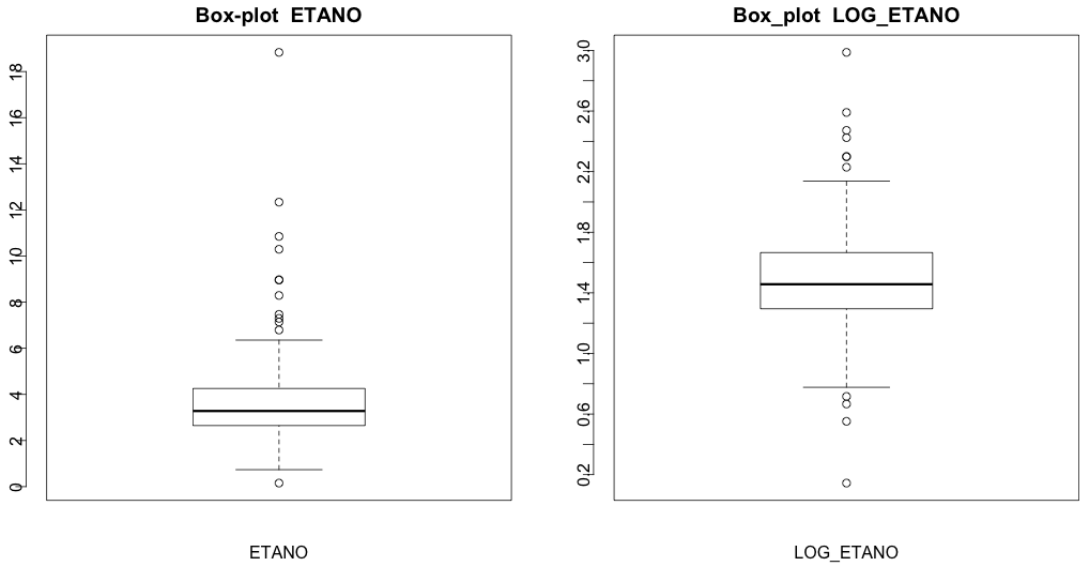


Fonte: Autor (2012)

Os gráficos Box-plot das variáveis: ETANO e LOG_ETANO são apresentados na Figura 20. Em ambos os gráficos é possível observar a presença de valores atípicos, além disso, fica evidente que, através da transformação LOG_ETANO

(direita), as estatísticas para média e mediana apresentam valores mais centrados ao conjunto de dados do que em relação a variável original ETANO.

Figura 20 - Box-plot para ao dados da variável ETANO (à esquerda) e para os dados da variável LOG_ETANO (à direita).



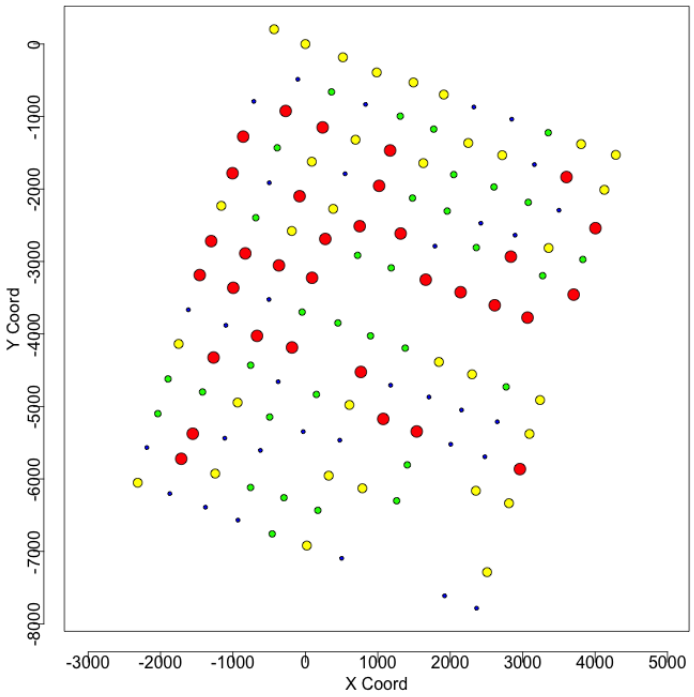
Fonte: Autor (2012)

Será dado seguimento às análises utilizando apenas a variável LOG_ETANO por apresentar melhores características em sua distribuição.

4.2.2 Análise Exploratória Espacial

A função *points()* do pacote *geoR* foi a ferramenta utilizada para gerar o gráfico da localização geográfica da variável LOG_ETANO (Figura 21). As cores e tamanho dos pontos (círculos) plotados no gráfico representam cada quartil do conjunto de dados. Círculos menores e de cor azul correspondem ao primeiro quartil, seguindo pela cor verde, amarelo e vermelho que se referem ao segundo, terceiro e quarto quartis, respectivamente. Na Figura 21 é possível visualizar a localização geográfica da variável LOG_ETANO. Graficamente, observa-se uma clara concentração de valores elevados na área central da figura e de valores menores nas extremidades superior e inferior.

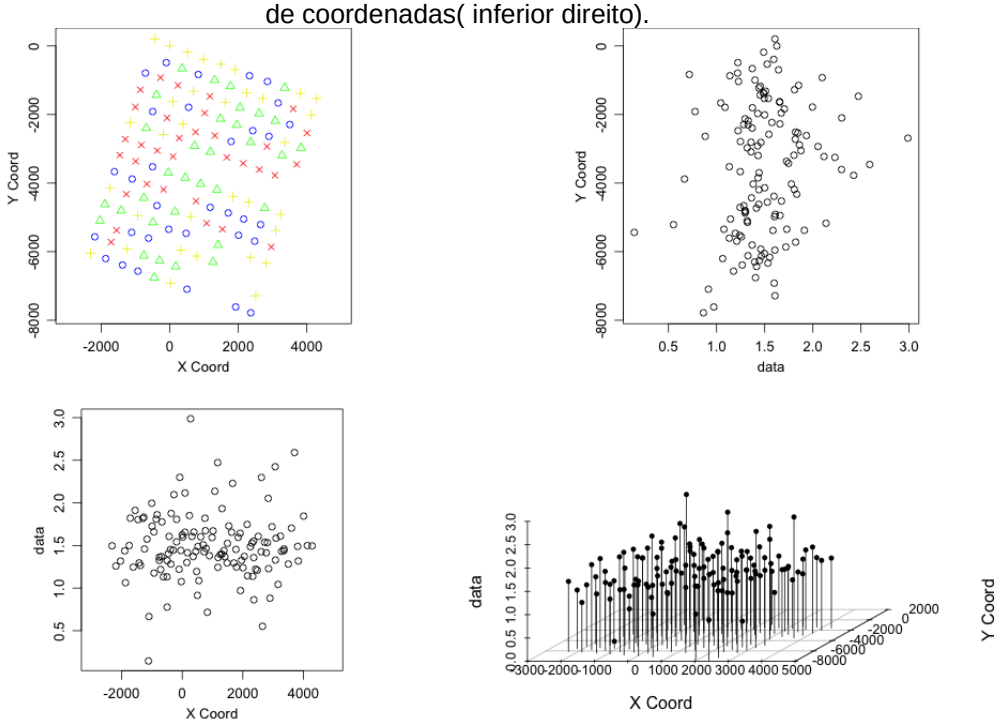
Figura 21 - Localização geográfica do LOG_ETANO associando cores fortes com a magnitude dos dados.



Fonte: Autor (2012)

Observa-se graficamente na Figura 22 a não existência de tendência linear na dependência espacial associada às coordenadas da variável LOG_ETANO. O comportamento dos dados quando cruzados com os eixos X e Y não apresenta qualquer forma evidente. O comando *plot.geodata()* foi o método utilizado para gerar a Figura 22 resultando em um resumo descritivo exploratório espacial.

Figura 22 - Localização geográfica do LOG_ETANO (superior esquerdo), valores do LOG_ETANO versus as coordenadas (superior direito e inferior esquerdo), e valores do LOG_ETANO sobre o plano de coordenadas(inferior direito).



Fonte: Autor (2012)

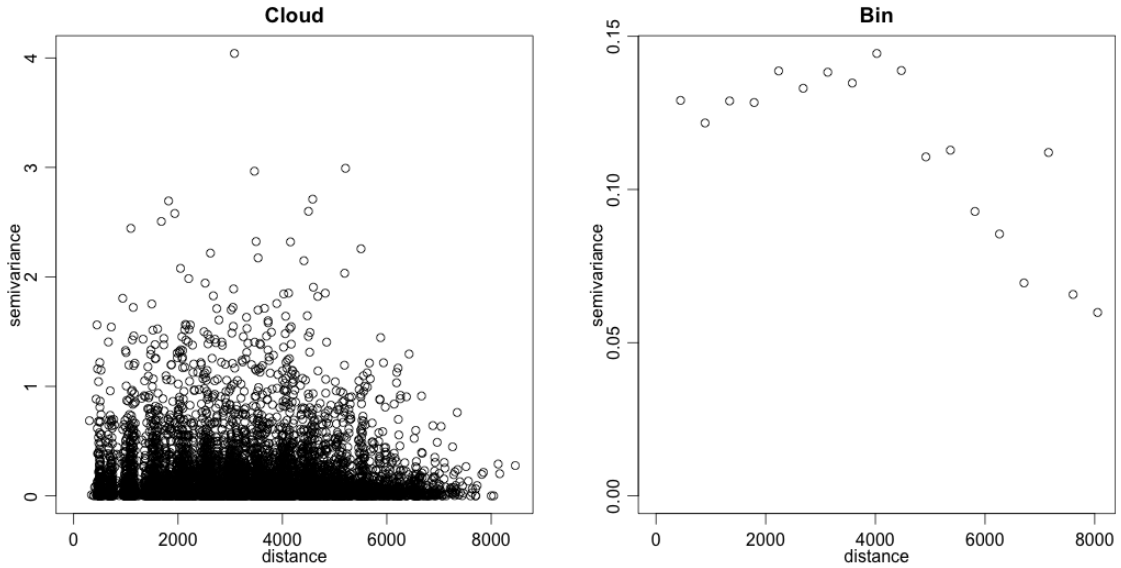
4.2.3 Análise Geoestatística

Assim como na metodologia utilizada na análise geoestatística do metano, na investigação da dependência espacial da variável LOG_ETANO é de interesse encontrar características que auxiliem na modelagem da distribuição espacial, além de verificar a associação do gás com os fatores físicos de Umidade, Tipo de Solo, Cor e Uso do Solo.

Novamente, o primeiro passo da análise geoestatística deu-se através do uso dos variogramas empíricos de nuvem (*cloud*) e pontuais (*bins*) (Figura 23). Analisando o gráfico do variograma *cloud* pode-se verificar que *outliers* possivelmente estão superestimando as semivariâncias em diferentes distâncias entre pares. Visualmente, este fenômeno está também afetando as menores distâncias e pode ser verificado pela alta dispersão dos pontos do gráfico do variograma *cloud* nesta região. Através do variograma *bin* pode-se comprovar esta verificação no momento em que o valor do *bin*, já no primeiro *lag*, praticamente alcança o patamar do variograma. Ressalta-se que os *bins* foram calculados

utilizando o estimador robusto (modulus) e para todo o conjunto de possíveis distâncias entre pares.

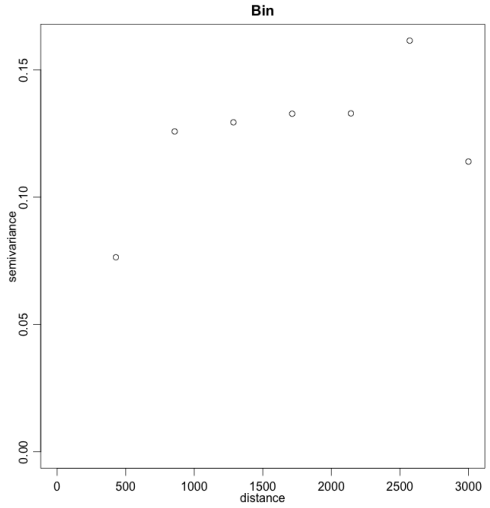
Figura 23 – Variograma Cloud Omnidirecional(esquerda) e o variograma Bin Omnidirecional (direita) da variável LOG_ETANO.



Fonte: Autor (2012)

O variograma *bin* apresentado na Figura 23 não apresentou um formato da distribuição do pontos que justificasse a escolha de um modelo para representar o variograma. Porém, pode-se verificar na análise descritiva espacial que visualmente era clara a presença de dependência no processo devido ao mesmo apresentar regiões bem definidas quando avaliadas através dos quartis. Os apontamentos realizados anteriormente sobre os *outliers* sugerem a necessidade de uma investigação mais detalhada nas menores distâncias. Diferentes configurações da distância máxima e de número de *lags* foram utilizadas como parâmetros da função *variog()* do pacote *geoR* para gerar variogramas durante esta investigação. Buscava-se encontrar algum variograma que apresentasse um formato que auxiliasse na escolha de um modelo que os representasse. Dentre os variogramas investigados, a Figura 24 mostra o que apresentou em seu formato características que mais se assemelham às dos modelos paramétricos conhecidos. As configurações utilizadas foram: distância máxima 3000 com 8 *lags*.

Figura 24 - Variograma Bin com distância máxima de 3000.



Fonte: Autor (2012)

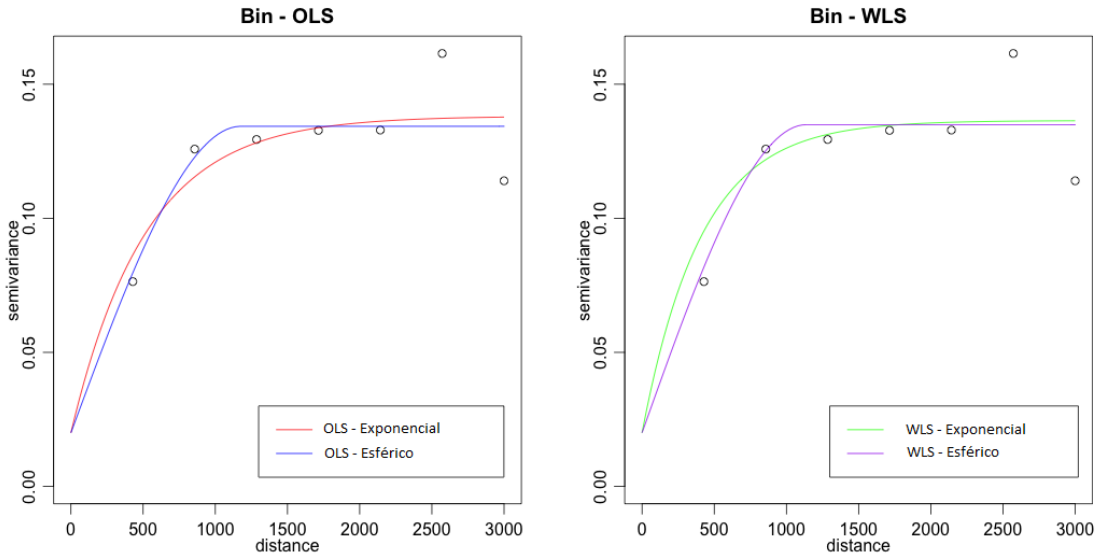
Novamente, o formato gráfico do variograma justifica a escolha por modelos exponencial e esférico para representar a estrutura de covariância, desta vez ajustados aos pontos do variograma dos dados da variável LOG_ETANO. Os métodos de mínimos quadrados ordinários (OLS) e mínimos quadrados ponderados (WLS) foram utilizados para estimação dos parâmetros do modelo. O parâmetro efeito pepita (*nugget*) foi fixado e os valores iniciais solicitados pela função *variofit()* foram os mesmos para os quatro modelos investigados. Os resultados obtidos são apresentados na Tabela 8 e os ajustes obtidos estão plotados na Figura 25.

Tabela 8 - Resultado das estimativas dos parâmetros para os dados do LOG_ETANO.

Estimadores	OLS		WLS	
Parâmetros	Exponencial	Esférico	Exponencial	Esférico
Nugget	0,0200	0,0200	0,0200	0,0150
Sill	0,1181	0,1143	0,1165	0,1149
Range	519,0801	1.177,0587	412,4150	1.138,8139

Fonte: Autor (2012)

Figura 25 - Variogramas ajustados utilizando os métodos de mínimos quadrados ordinários (OLS) e mínimos quadrados ponderados (WLS) pelos modelos exponencial e esférico.



Fonte: Autor (2012)

No esforço de modelar os dados utilizando os métodos de máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) com os modelos exponencial e esférico através da função likfit() algumas restrições foram necessárias para obtenção de resultados aceitáveis.

Ao aplicar os modelos exponenciais, tanto por ML como por REML, observou-se que ao não fixar o efeito pepita (*nugget*) na modelagem, o mesmo acabava sendo superestimado pelo método, consequentemente subestimando o parâmetro *sigmasq* e gerando estimativas elevadas para *phi*. Com isso, as superfícies preditas por esse modelo eram muito suavizadas e com predições apenas em torno da média geral do fenômeno. Ao fixar o efeito pepita (*nugget=0,02*) as estimativas para *phi* passaram a resultar valores menores que a menor distância entre pares. Optou-se então por delimitar o intervalo de possíveis estimativas para *phi* através do argumento *limits=pars.limits(phi=c(520,1000))* onde finalmente conseguiu-se um modelo que se adaptasse melhor a superfície de dados.

As mesmas dificuldades que ocorreram durante uso do modelo exponencial, ocorreram novamente com o modelo esférico, muito possivelmente devido aos *outliers* observados. Voltou-se a optar por delimitar o intervalo de *phi*, desta vez com valores entre 1100 e 1200 (por ser um intervalo próximo aos valores estimados por OLS e WLS).

Foram obtidos os valores para as estimativas dos parâmetros, AIC e logaritmo da função de verossimilhanças utilizando o estimador de máxima verossimilhança com estrutura de covariância exponencial nos dados da variável LOG_ETANO em modelos sem covariável(constante) e com cada uma das covariáveis: Umidade, Tipo de Solo, Cor e Uso do Solo (Tabela 9).

Os resultados do teste da razão de verossimilhanças (Tabela 9), quando comparado o modelo com a covariável e o modelo sem a covariável (constante), foram menores que o valor crítico 5,99 ($\chi^2_{(2;0,05)}$). Portanto, ao nível de significância de 5%, conclui-se que os modelos com covariável não foram significativamente melhores que o modelo 21. Porém, o critério de informação de Akaike (AIC) leva a uma decisão contraditória quando o modelo 23 apresenta menor valor AIC. Alguns estudos acadêmicos colocam que, para a diferença ser significativa a diferença no valor do critério deve ser maior que 3. Como o teste da razão de verossimilhanças não apontou diferença significativa, dá-se seguimento às análises selecionando o modelo 21. Vale destacar, que ao limitar o intervalo de possíveis valores para *phi*, o método estimou os parâmetros, em todos os modelos, o valor inferior deste intervalo.

Tabela 9 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_ETANO, com estrutura de covariância exponencial, utilizando os estimadores de máxima verossimilhança.

	Modelo 21	Modelo 22	Modelo 23	Modelo 24	Modelo 25
Método	ML	ML	ML	ML	ML
Modelo ajuste	Exponencial	Exponencial	Exponencial	Exponencial	Exponencial
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,02	0,02	0,02	0,02	0,02
PartialSill	0,1735	0,169	0,1666	0,1649	0,1717
Phi	520	520	520	520	520
AIC	141,1798	141,9446	140,1919	141,7872	147,9362
LogL	-67,5899	-65,9723	-65,096	-64,8936	-66,9681
Razão de verossimilhança		3,2352	4,9878	5,3926	1,2436
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

Utilizando o estimador de máxima verossimilhança restrita com estrutura de covariância exponencial (Tabela 10), novamente o método estimou o valor informado no limite inferior do intervalo para possíveis valores de *phi* que foi utilizado na função *likfit()*. Avaliando os testes da razão de verossimilhanças, o modelo 29 foi o único a apresentar valor superior a 5,99, que significa que, ao nível de significância de 5%, o ajuste deste modelo foi significativamente melhor do que o modelo 26, o qual se refere ao modelo sem covariável. O menor valor de AIC foi o do modelo 28, porém quando comparado aos outros modelos esta diferença não foi superior a 3. Portanto, o modelo 29 foi selecionado.

Tabela 10 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_ETANO, com estrutura de covariância exponencial, utilizando os estimadores de máxima verossimilhança restrita.

	Modelo 26	Modelo 27	Modelo 28	Modelo 29	Modelo 30
Método	REML	REML	REML	REML	REML
Modelo ajuste	Exponencial	Exponencial	Exponencial	Exponencial	Exponencial
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,02	0,02	0,02	0,02	0,02
PartialSill	0,1754	0,1741	0,1716	0,1716	0,1802
Phi	520	520	520	520	520
AIC	139,5182	139,8272	138,2675	139,2919	146,2269
LogL	-66,7591	-64,9136	-64,1337	-63,6459	-66,1134
Razão de verossimilhança		3,691	5,2508	6,2264	1,2914
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

Os resultados apresentados na Tabela 11 e Tabela 12 referem-se ajustes utilizando estimadores de máxima verossimilhança e máxima verossimilhança restrita, respectivamente, com estrutura de covariância esférica. É importante colocar que, ao limitar o intervalo de possíveis valores para *phi*, ambos os métodos estimaram o parâmetro, em todos os modelos ajustados, com o valor inferior do intervalo informado na função *likfit()*. Tanto pelo método ML, como o método REML, os resultados para os testes da razão de verossimilhanças apontam para os modelos com a covariável Tipo de Solo. É possível observar que os modelos 33 e 38

apresentam valor superior a 5,99, que significa que, ao nível de significância de 5%, o ajuste destes modelos foi significativamente melhor do que o modelo 31 e 36, respectivamente.

Tabela 11 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_ETANO, com estrutura de covariância esférica, utilizando os estimadores de máxima verossimilhança.

	Modelo 31	Modelo 32	Modelo 33	Modelo 34	Modelo 35
Método	ML	ML	ML	ML	ML
Modelo ajuste	Esférico	Esférico	Esférico	Esférico	Esférico
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,02	0,02	0,02	0,02	0,02
PartialSill	0,1856	0,1794	0,1768	0,1768	0,1823
Phi	1100	1100	1100	1100	1100
AIC	152,7174	152,5192	150,5782	153,7617	158,6055
LogL	-73,3587	-71,2596	-70,2891	-70,8809	-72,3028
Razão de verossimilhança		4,1982	6,1392	4,9556	2,1118
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

Tabela 12 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do LOG_ETANO, com estrutura de covariância esférica, utilizando os estimadores de máxima verossimilhança restrita.

	Modelo 36	Modelo 37	Modelo 38	Modelo 39	Modelo 40
Método	REML	REML	REML	REML	REML
Modelo ajuste	Esférico	Esférico	Esférico	Esférico	Esférico
Tendência	Constante	Umidade	Tipo de Solo	Cor	Uso do Solo
Nugget	0,02	0,02	0,02	0,02	0,02
PartialSill	0,1876	0,1851	0,1824	0,1844	0,1918
Phi	1100	1100	1100	1100	1100
AIC	151,557	150,8125	149,0337	151,6292	157,1546
LogL	-72,7785	-70,4062	-69,5169	-69,8146	-71,5773
Razão de verossimilhança		4,7446	6,5232	5,9278	2,4024
Num. De Parâmetros	3	5	5	6	7

Fonte: Autor (2012)

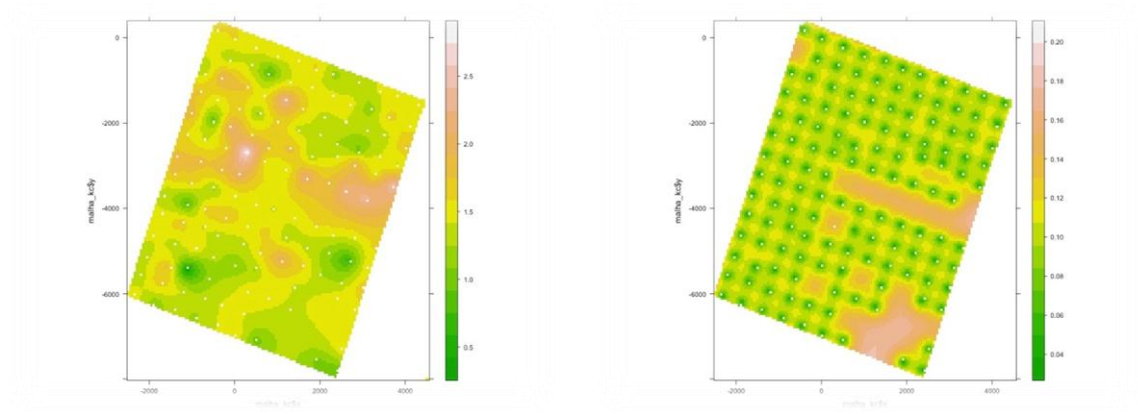
4.2.4 Predições

Uma vez estimados os parâmetros para os modelos, com suas respectivas estruturas de covariância, através dos métodos ML e REML, é realizada a krigagem sobre os modelos selecionados para fazer a predição da superfície em locais não observados. A krigagem ordinária foi o método empregado para obter os resultados das predições e variância das predições do LOG_ETANO.

Os resultados das predições e variância das predições dos modelos para a variável LOG_ETANO, para cada um dos quatro modelos selecionados, estão apresentados abaixo: (Figura 26, Figura 27, Figura 28 e Figura 29)

Figura 26 - Resultados da predição (a) e variância da predição (b) para o LOG_ETANO, referentes ao ajuste do modelo 21.

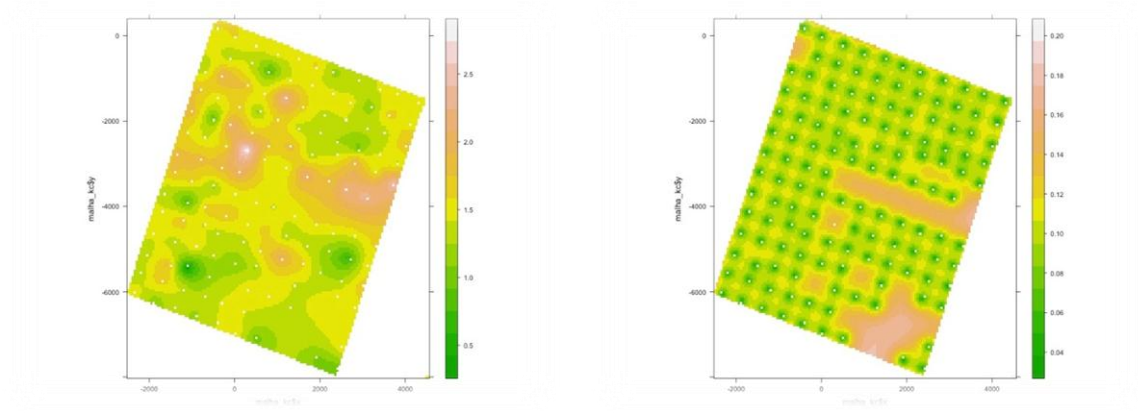
a) b)



Fonte: Autor (2012)

Figura 27 - Resultados da predição (a) e variância da predição (b) para o LOG_ETANO, referentes ao ajuste do modelo 29.

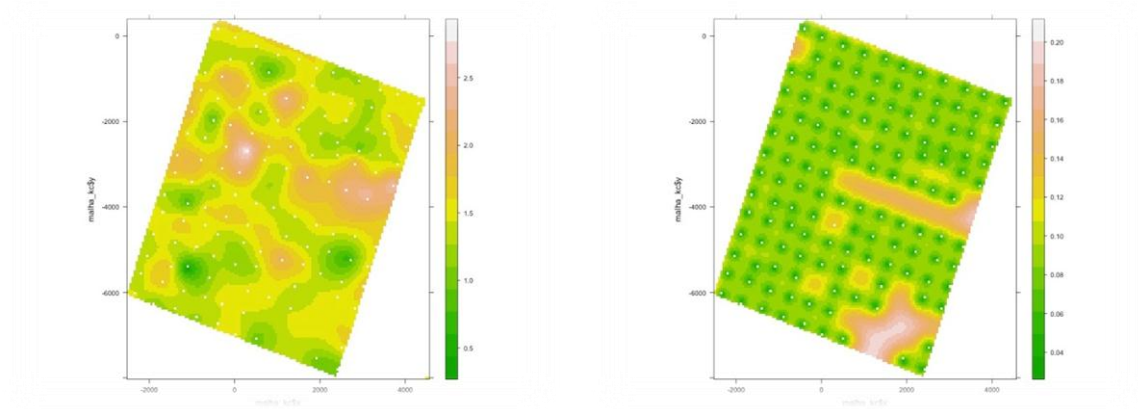
a) b)



Fonte: Autor (2012)

Figura 28 - Resultados da predição (a) e variância da predição (b) para o LOG_ETANO, referentes ao ajuste do modelo 33.

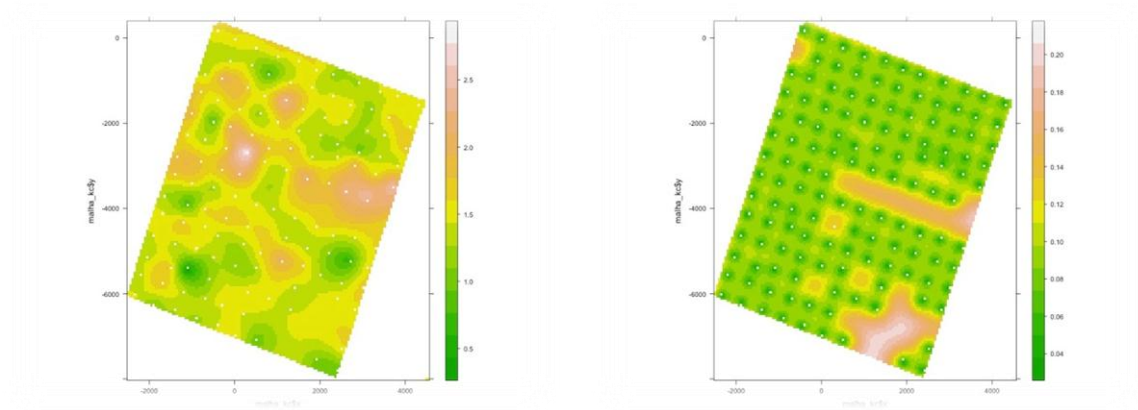
a) b)



Fonte: Autor (2012)

Figura 29 - Resultados da predição (a) e variância da predição (b) para o LOG_ETANO, referentes ao ajuste do modelo 38.

a) b)



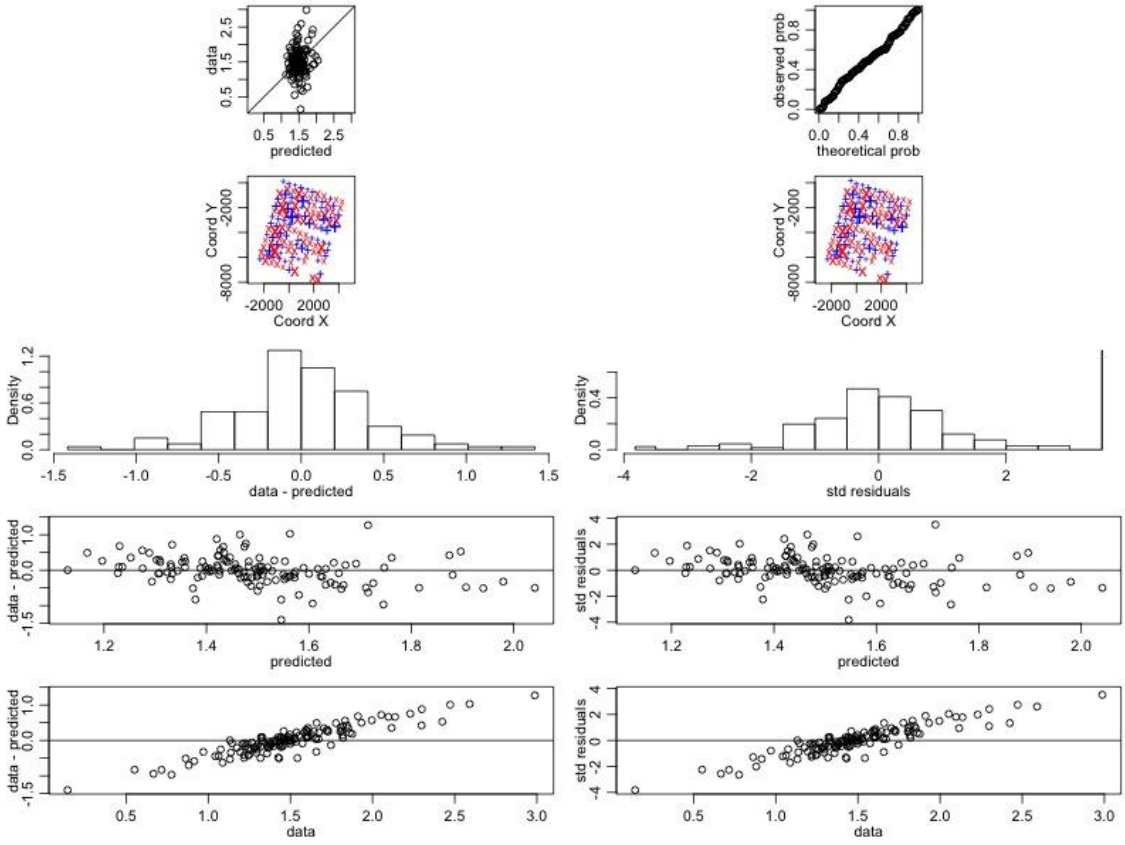
Fonte: Autor (2012)

Constata-se que os mapas gerados pelos métodos ML e REML com estrutura de covariância exponencial apresentaram superfícies muito semelhantes. Os modelos esféricos ML e REML também apresentaram superfícies bastante parecidas. Todos os modelos resultaram variâncias associadas às predições muito próximas, porém foi possível verificar, através da validação cruzada, que o modelo 38 consegue captar com mais detalhes a amplitude dos dados amostrados do LOG_ETANO. Portanto, acredita-se que este modelo é o que melhor representa a estrutura espacial dos dados do LOG_ETANO.

4.2.5 Validação cruzada

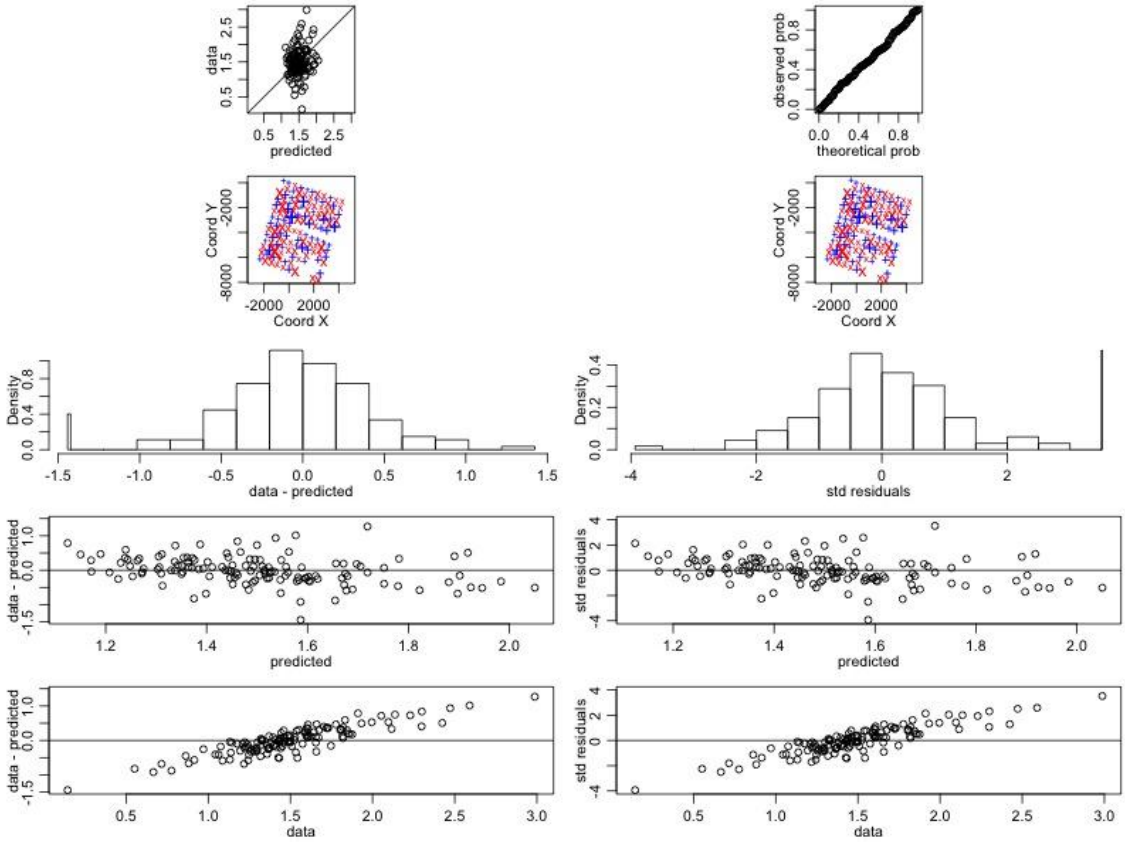
A Figura 30, Figura 31, Figura 32 e Figura 33 apresentam a validação cruzada para a variável LOG_ETANO. Os resultados obtidos foram satisfatórios para todos os modelos selecionados (21,29,33,38), e esta diretamente associado à qualidade do modelo escolhido e das estimativas obtidas. Verificou-se que a distribuição dos erros positivos e negativos está dispersa pela região de estudo; os valores preditos estão próximo a reta padrão, com exceção do pontos associados a outliers e os dados da probabilidade teórica e observada se encontram sobre a reta, indicando que a predição foi eficiente através do uso do método da krigagem ordinária.

Figura 30 - Validação cruzada para o LOG_ETANO referente ao modelo 21.



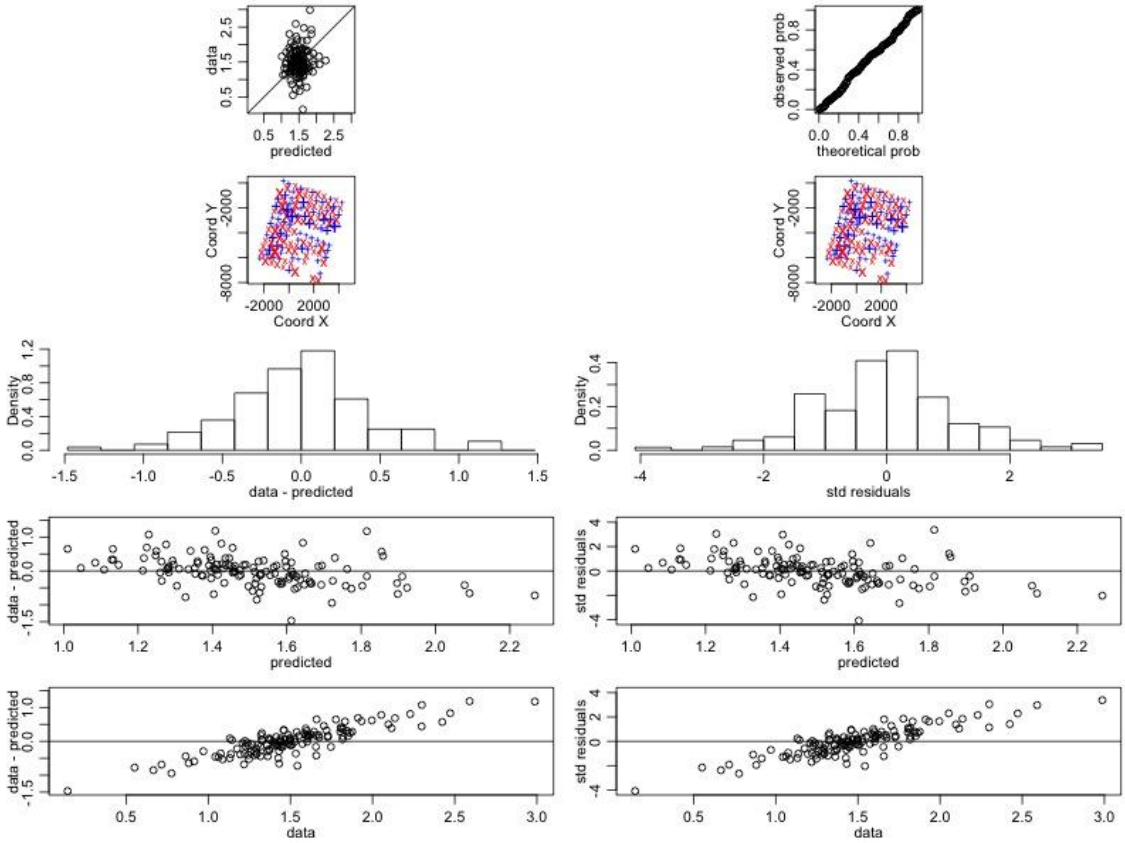
Fonte: Autor (2012)

Figura 31 - Validação cruzada para o LOG_ETANO referente ao modelo 29.



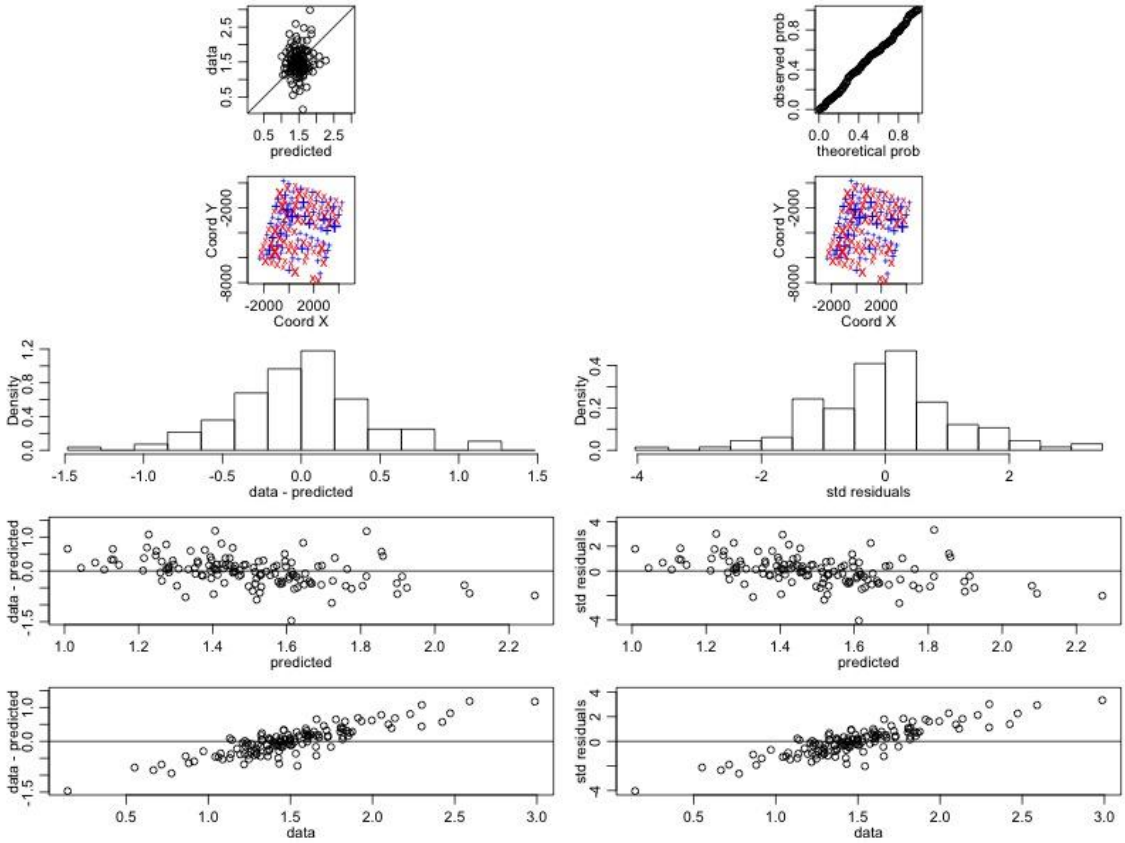
Fonte: Autor (2012)

Figura 32 - Validação cruzada para o LOG_ETANO referente ao modelo 33.



Fonte: Autor (2012)

Figura 33 - Validação cruzada para o LOG_ETANO referente ao modelo 38.



Fonte: Autor (2012)

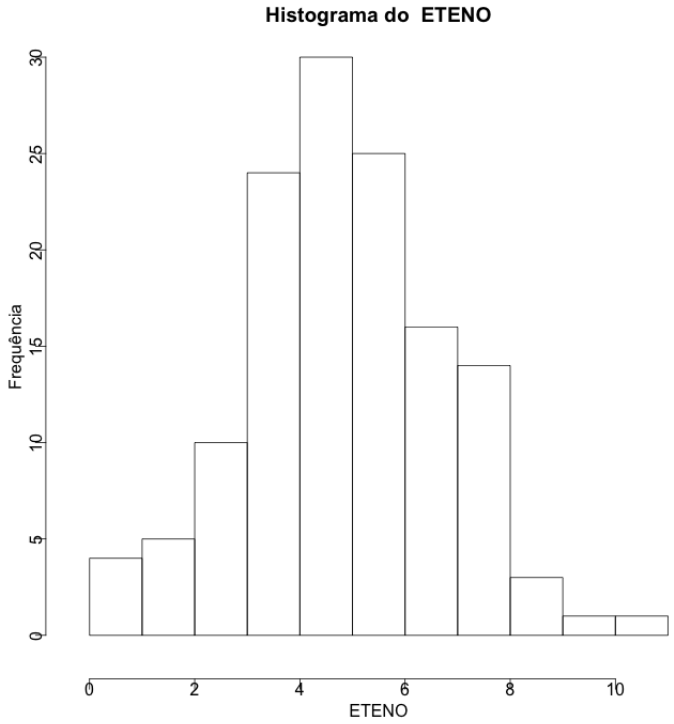
4.3 ANÁLISE DA VARIÁVEL ETENO

Os mesmos procedimentos de análise utilizados anteriormente serão utilizados neste capítulo para o estudo da variável ETENO.

4.3.1 Análise Descritiva

Diferentemente dos gases metano e etano, os dados referentes à variável ETENO apresenta uma distribuição claramente simétrica conforme histograma apresentado na Figura 34. Portanto, para realização desta análise não se faz necessária qualquer transformação nos dados originais.

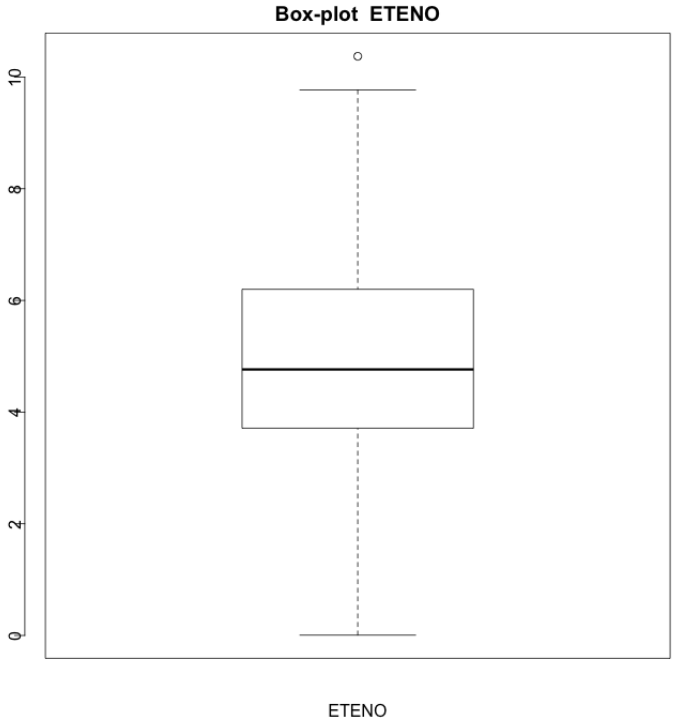
Figura 34 - Representação da distribuição dos dados originais do ETANO.



Fonte: Autor (2012)

A Figura 35 apresenta o gráfico Box-plot da variável ETENO. É possível observar no gráfico que os valores para média e mediana apresentam-se centrados ao conjunto de dados, além disso, apenas um *outlier* é verificado no gráfico.

Figura 35 - Box-plot para ao dados da variável ETENO.

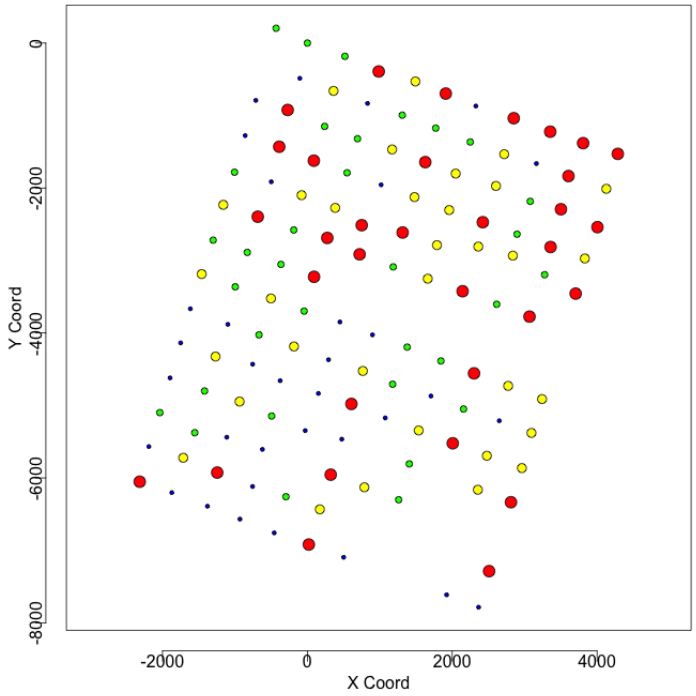


Fonte: Autor (2012)

4.3.2 Análise Exploratória Espacial

O gráfico de círculos que apresenta a localização geográfica destacando cada quartil do conjunto de dados da variável ETENO é apresentado na Figura 36. No gráfico é possível observar concentrações de cores semelhantes que apontam, visualmente, claros indícios de dependência espacial.

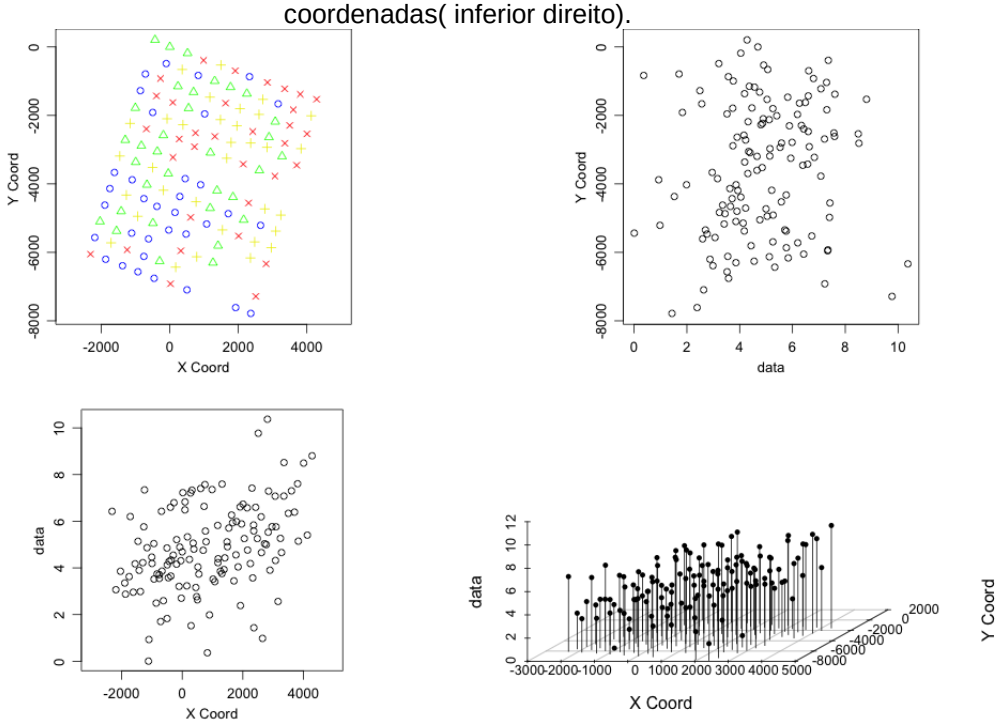
Figura 36 - Localização geográfica do ETENO associando cores fortes com a magnitude dos dados.



Fonte: Autor (2012)

Analisando os gráficos referentes ao valor da variável ETENO versus os eixos de coordenadas X e Y plotados na Figura 37 (inferior esquerdo e superior direito, respectivamente) observa-se que existe uma provável tendência linear ou quadrática associada à coordenada X sobre a dependência espacial do ETENO. Posteriormente, quando forem ajustados os modelos, será realizada a verificação desta associação.

Figura 37 - Localização geográfica do ETENO (superior esquerdo), valores do ETENO versus as coordenadas (superior direito e inferior esquerdo), e valores do ETENO sobre o plano de coordenadas(inferior direito).



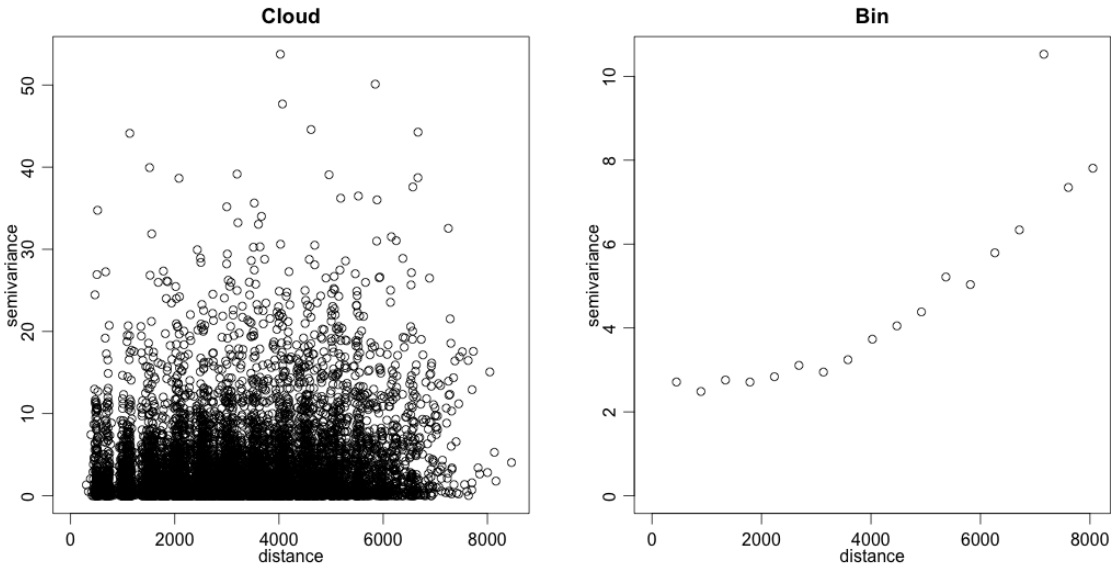
Fonte: Autor (2012)

4.3.3 Análise Geoestatística

Novamente, a mesma metodologia de análise geoestatística será utilizada. Portanto, o primeiro passo tem como objetivo encontrar características que auxiliem na modelagem da dependência espacial da variável ETENO. Relembrando que também é de interesse verificar a associação do gás com os mesmos fatores físicos de Umidade, Tipo de Solo, Cor e Uso do Solo.

O primeiro passo da análise geoestatística da variável ETENO foi através da utilização dos variogramas empíricos de nuvem (*cloud*) e pontuais (*bins*) (Figura 38). O variograma *cloud* mostra a existência de uma acentuada dispersão nas semivariâncias. Verifica-se também que esta dispersão aumenta a partir da distância 4000. O variograma *Bin* confirma este aumento na dispersão com o claro aumento no valor médio das semivariâncias a partir da distância 4000. Foi utilizado o estimador robusto (*modulus*) para calcular o variograma *Bin*. Já o variograma *cloud* foi calculado através do método clássico.

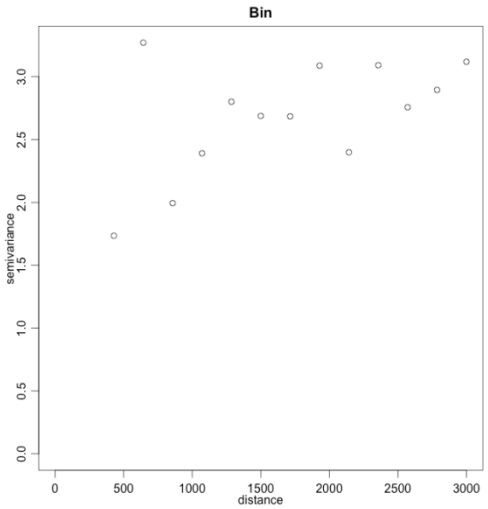
Figura 38 – Variograma Cloud Omnidirecional(esquerda) e o variograma Bin Omnidirecional (direita) da variável ETENO.



Fonte: Autor (2012)

Foi realizada uma investigação mais detalhada em distâncias menores para auxiliar na escolha de um modelo para representar o variograma. Diferentes configurações da distância máxima e de número de *lags* foram utilizadas como parâmetros da função *variog()* do pacote *geoR* durante a investigação. Optou-se por realizar a modelagem sobreo variograma configurado com os argumentos distância máxima igual a 3000 e o número de *lags* definido como 15. A Figura 39 mostra o gráfico do variograma selecionado.

Figura 39 - Variograma Bin com distância máxima de 3000.



Fonte: Autor (2012)

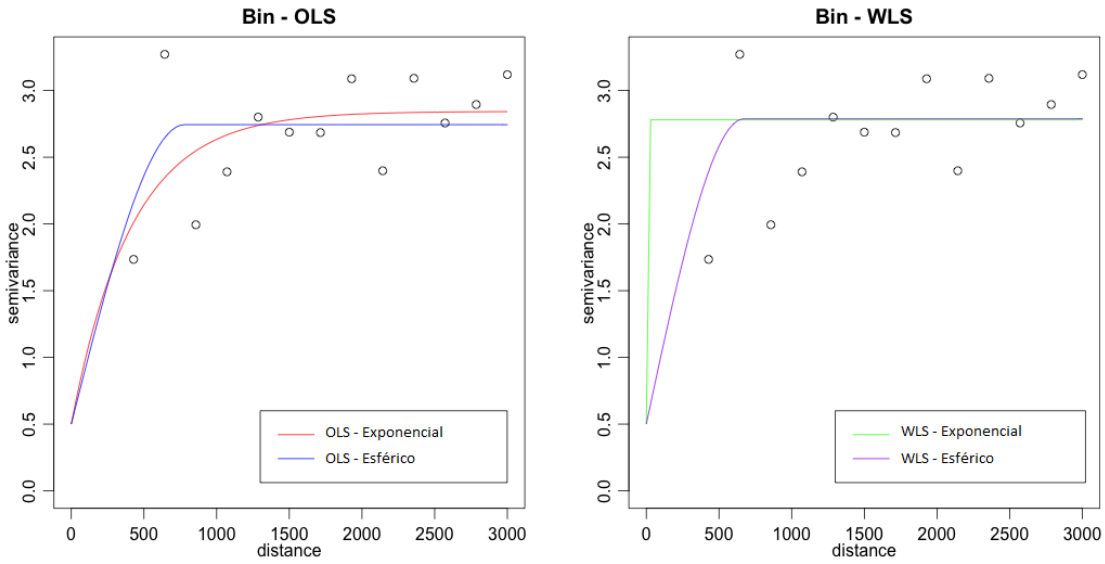
Foram novamente escolhidos os modelos exponencial e esférico para representar a estrutura de covariância, agora ajustados aos pontos do variograma dos dados da variável ETENO. Os métodos de mínimos quadrados ordinários (OLS) e mínimos quadrados ponderados (WLS) foram utilizados para estimação dos parâmetros do modelo. O efeito pepita (*nugget*) foi fixado com o valor 0,5 e a distância máxima foi fixada com o valor 3000 nos quatro modelos ajustados pela função *variofit()*. Os resultados obtidos são apresentados na Tabela 13 e os ajustes obtidos estão plotados na Figura 40.

Tabela 13 - Resultado das estimativas dos parâmetros para os dados do ETENO.

Estimadores	OLS		WLS	
Parâmetros	Exponencial	Esférico	Exponencial	Esférico
Nugget	0,5	0,5	0,5	0,5
Sill	2,3432	2,2441	2,2802	2,2880
Range	413,9409	779,6092	0	671,7925

Fonte: Autor (2012)

Figura 40 - Variogramas ajustados utilizando os métodos de mínimos quadrados ordinários (OLS) e mínimos quadrados ponderados (WLS) pelos modelos exponencial e esférico.



Fonte: Autor (2012)

Ao realizar a modelagem através dos métodos de máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) com os modelos exponencial e esférico novamente foram necessárias algumas intervenções, por meio de

argumentos da função likfit() para que fosse possível encontrar resultados adequados.

Em ambos os modelos, exponencial e esférico, foi fixado o efeito pepita e delimitado o intervalo de possíveis estimativas para o *phi* para a utilização dos métodos ML e REML. É importante salientar que quando não fixados estes parâmetros, o método não conseguiu chegar a bons resultados, porém vale destacar que este tipo de manipulação deve ser utilizado após testar a modelagem sem estas restrições.

Primeiramente, foi realizado o ajuste para verificar a associação do vetor de coordenadas na estrutura espacial dos dados da variável ETENO. A covariável do vetor de coordenadas X apresentou associação significativa, portanto os modelos testados neste estudo da estrutura espacial do ETENO irão incorporar esta covariável.

Foram obtidos os valores para as estimativas dos parâmetros, AIC e logaritmo da função de verossimilhanças utilizando o estimador de máxima verossimilhança com estrutura de covariância exponencial nos dados da variável ETENO em modelos sem covariável(constante) e com cada uma das covariáveis: Umidade, Tipo de Solo, Cor e Uso do Solo (Tabela 14).

Os resultados do teste da razão de verossimilhanças (Tabela 14), quando comparado o modelo com a covariável mais a tendência linear nas coordenadas e o modelo somente com a tendência linear nas coordenadas, apresentaram valores maiores que o valor crítico 5,99 ($\chi^2_{(2;0,05)}$) para os modelos 42, 43 e 44. Portanto, ao nível de significância de 5%, conclui-se que estes modelos são significativamente melhores que o modelo 41. O critério de informação de Akaike (AIC) apresentou menor valor no modelo 43, portanto este será o modelo. Vale destacar, que ao limitar o intervalo de possíveis valores para *phi*, o método estimou os parâmetros, em todos os modelos, o valor inferior deste intervalo.

Tabela 14 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do ETENO, com estrutura de covariância exponencial, utilizando os estimadores de máxima verossimilhança.

	Modelo 41	Modelo 42	Modelo 43	Modelo 44	Modelo 45
Método	ML	ML	ML	ML	ML
Modelo ajuste	Exponencial	Exponencial	Exponencial	Exponencial	Exponencial
Tendência	Coordenadas	Coordenadas + Umidade	Coordenadas + Tipo de Solo	Coordenadas + Cor	Coordenadas + Uso do Solo
Nugget	0,5	0,5	0,5	0,5	0,5
PartialSill	3,755	3,4814	3,4491	3,4981	3,581
Phi	600	600	600	600	600
AIC	547,8352	543,3698	542,3339	546,6941	550,5305
LogL	-268,9176	-264,6849	-264,167	-265,3471	-266,2652
Razão de verossimilhança		8,4654	9,5012	7,141	5,3048
Num. De Parâmetros	5	7	7	8	9

Fonte: Autor (2012)

Utilizando o estimador de máxima verossimilhança restrita com estrutura de covariância exponencial (Tabela 15), novamente o método estimou o valor informado no limite inferior do intervalo para possíveis valores de *phi* que foi utilizado na função *likfit()*. Analisando os testes da razão de verossimilhanças, pode-se verificar que todos os modelos com as covariáveis referentes aos fatores físicos, quando adicionados estes fatores no modelo que tem como parâmetro apenas a tendência linear nas coordenadas, contribuem significativamente, ao nível de significância de 5%, na explicação do fenômeno da distribuição espacial da variável ETENO. Isso é observado devido aos modelos 47, 48, 49 e 50 apresentarem valor para a razão de verossimilhança superior a 5,99. O modelo que apresentou o menor valor de AIC foi o do modelo 48, portanto será o selecionado.

Tabela 15 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do ETENO, com estrutura de covariância exponencial, utilizando os estimadores de máxima verossimilhança restrita.

	Modelo 46	Modelo 47	Modelo 48	Modelo 49	Modelo 50
Método	REML	REML	REML	REML	REML
Modelo ajuste	Exponencial	Exponencial	Exponencial	Exponencial	Exponencial
Tendência	Coordenadas	Coordenadas + Umidade	Coordenadas + Tipo de Solo	Coordenadas + Cor	Coordenadas + Uso do Solo
Nugget	0,5	0,5	0,5	0,5	0,5
PartialSill	3,8882	3,6822	3,648	3,7418	3,8623
Phi	600	600	600	600	600
AIC	533,6139	522,983	522,1308	522,8477	524,5446
LogL	-261,807	-254,4915	-254,0654	-253,4239	-253,2723
Razão de verossimilhança		14,631	15,4832	16,7662	17,0694
Num. De Parâmetros	5	7	7	8	9

Fonte: Autor (2012)

O teste da razão de verossimilhanças (Tabela 16) refere-se aos resultados obtidos pelo estimador de máxima verossimilhança com estrutura de covariância esférica quando compara um modelo com a covariável mais a tendência linear nas coordenadas e o modelo somente com a tendência linear nas coordenadas. Os modelos 52, 53 e 54 apresentaram valores maiores que o valor crítico 5,99 ($\chi^2_{(2;0,05)}$). Portanto, ao nível de significância de 5%, conclui-se que estes modelos são significativamente melhores que o modelo 51. O critério de informação de Akaike (AIC) apresentou menor valor no modelo 53, portanto este será o modelo selecionado. Vale destacar, que ao limitar o intervalo de possíveis valores para *phi*, o método estimou os parâmetros, em todos os modelos, o valor inferior deste intervalo.

Tabela 16 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do ETENO, com estrutura de covariância esférica, utilizando os estimadores de máxima verossimilhança.

	Modelo 51	Modelo 52	Modelo 53	Modelo 54	Modelo 55
Método	ML	ML	ML	ML	ML
Modelo ajuste	Esférico	Esférico	Esférico	Esférico	Esférico
Tendência	Coordenadas	Coordenadas + Umidade	Coordenadas + Tipo de Solo	Coordenadas + Cor	Coordenadas + Uso do Solo
Nugget	0,5	0,5	0,5	0,5	0,5
PartialSill	3,5544	3,266	3,2357	3,315	3,3836
Phi	1100	1100	1100	1100	1100
AIC	553,4427	548,175	546,7904	552,8308	555,9766
LogL	-271,7213	-267,0875	-266,3952	-268,4154	-268,9883
Razão de verossimilhança		9,2676	10,6522	6,6118	5,466
Num. De Parâmetros	5	7	7	8	9

Fonte: Autor (2012)

Ao usar o estimador de máxima verossimilhança restrita com estrutura de covariância esférica (Tabela 17), novamente o método estimou o valor informado no limite inferior do intervalo para possíveis valores de *phi* que foi utilizado na função *likfit()*. Os modelos 47, 48, 49 e 50 apresentaram valor para a razão de verossimilhança superior a 5,99, portanto a adição de suas respectivas covariáveis ao modelo que já tem a covariável tendência linear nas coordenadas, contribui significativamente, ao nível de significância de 5%, na explicação do fenômeno da distribuição espacial da variável ETENO. O modelo que apresentou o menor valor de AIC foi o do modelo 58, portanto será o selecionado.

Tabela 17 - Resultado das estimativas dos parâmetros, critério de informação de Akaike (AIC) e logaritmo da função de verossimilhanças para os dados do ETENO, com estrutura de covariância esférica, utilizando os estimadores de máxima verossimilhança restrita.

	Modelo 56	Modelo 57	Modelo 58	Modelo 59	Modelo 60
Método	REML	REML	REML	REML	REML
Modelo ajuste	Esférico	Esférico	Esférico	Esférico	Esférico
Tendência	Coordenadas	Coordenadas + Umidade	Coordenadas + Tipo de Solo	Coordenadas + Cor	Coordenadas + Uso do Solo
Nugget	0,5	0,5	0,5	0,5	0,5
PartialSill	3,6848	3,4625	3,4292	3,5573	3,6597
Phi	1100	1100	1100	1100	1100
AIC	541,227	529,7643	528,5232	530,9295	531,7742
LogL	-265,6135	-257,8821	-257,2616	-257,4647	-256,8871
Razão de verossimilhança		15,4628	16,7038	16,2976	17,4528
Num. De Parâmetros	5	7	7	8	9

Fonte: Autor (2012)

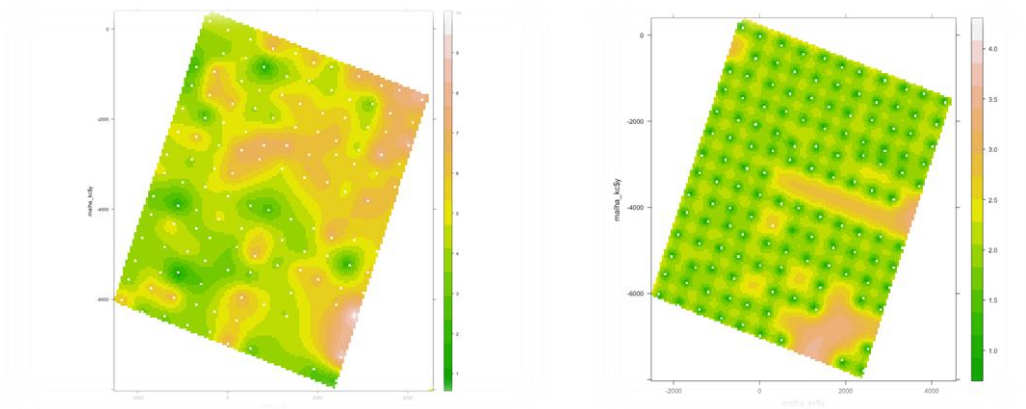
4.3.4 Predições

Estimados os parâmetros para os modelos através dos métodos ML e REML para às estruturas de covariância exponencial e esférica, é então realizada a krigagem sobre os modelos selecionados para fazer a predição da superfície em locais não observados. A krigagem ordinária foi o método empregado para obter os resultados das predições e variância das predições do ETENO.

Os resultados das predições e variância das predições dos modelos para a variável ETENO, para cada um dos quatro modelos selecionados, estão apresentados abaixo: (Figura 41, Figura 42, Figura 43 e Figura 44)

Figura 41 - Resultados da predição (a) e variância da predição (b) para o ETENO, referentes ao ajuste do modelo 43.

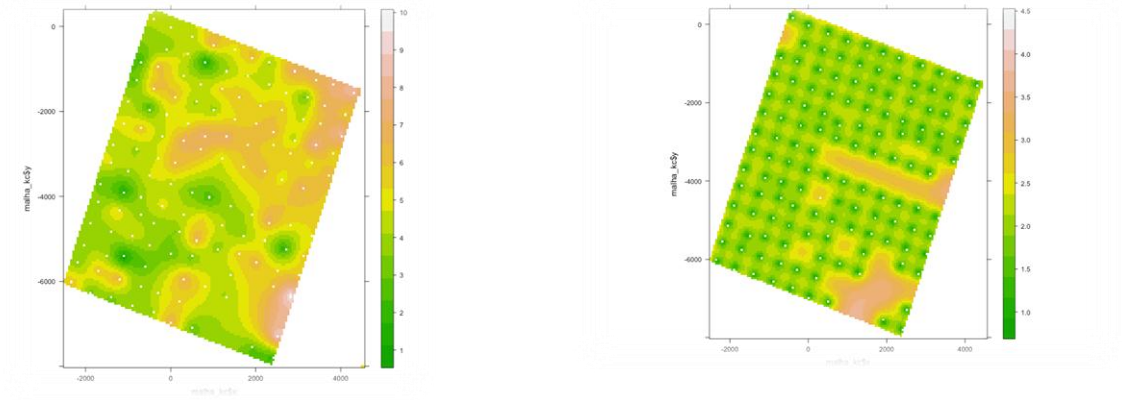
a) b)



Fonte: Autor (2012)

Figura 42 - Resultados da predição (a) e variância da predição (b) para o ETENO, referentes ao ajuste do modelo 48.

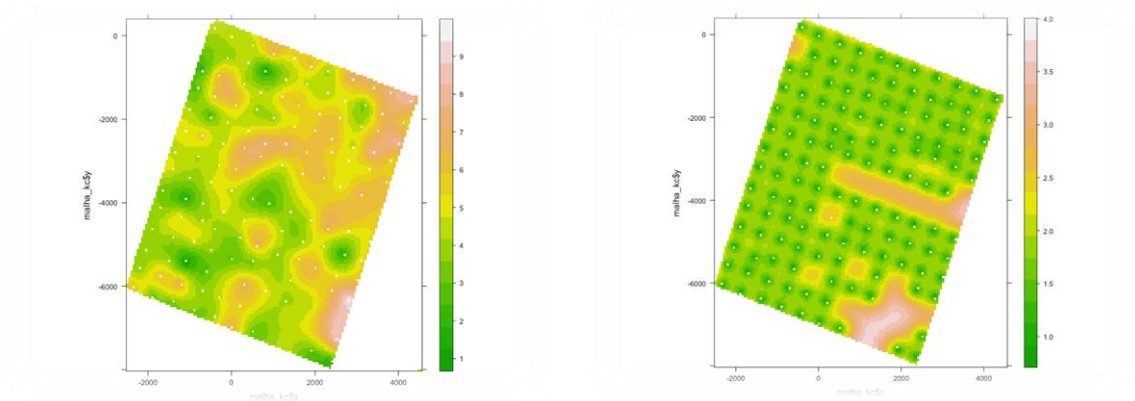
a) b)



Fonte: Autor (2012)

Figura 43 - Resultados da predição (a) e variância da predição (b) para o ETENO, referentes ao ajuste do modelo 53.

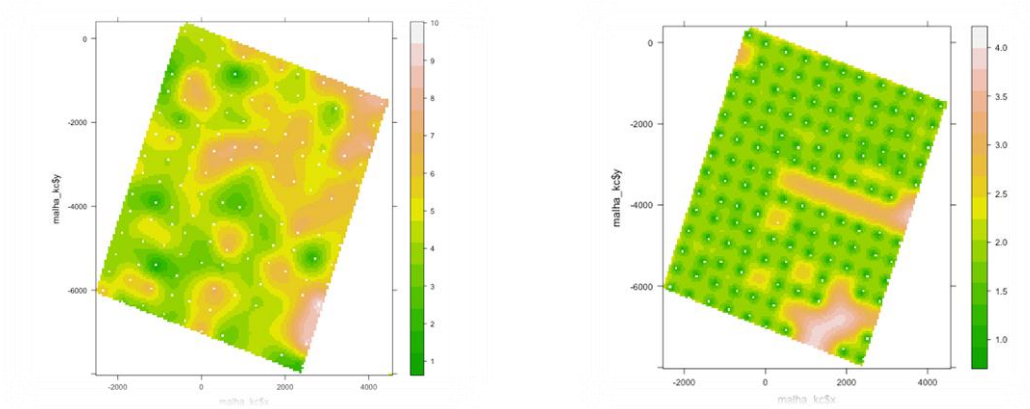
a) b)



Fonte: Autor (2012)

Figura 44 - Resultados da predição (a) e variância da predição (b) para o ETENO, referentes ao ajuste do modelo 58.

a) b)



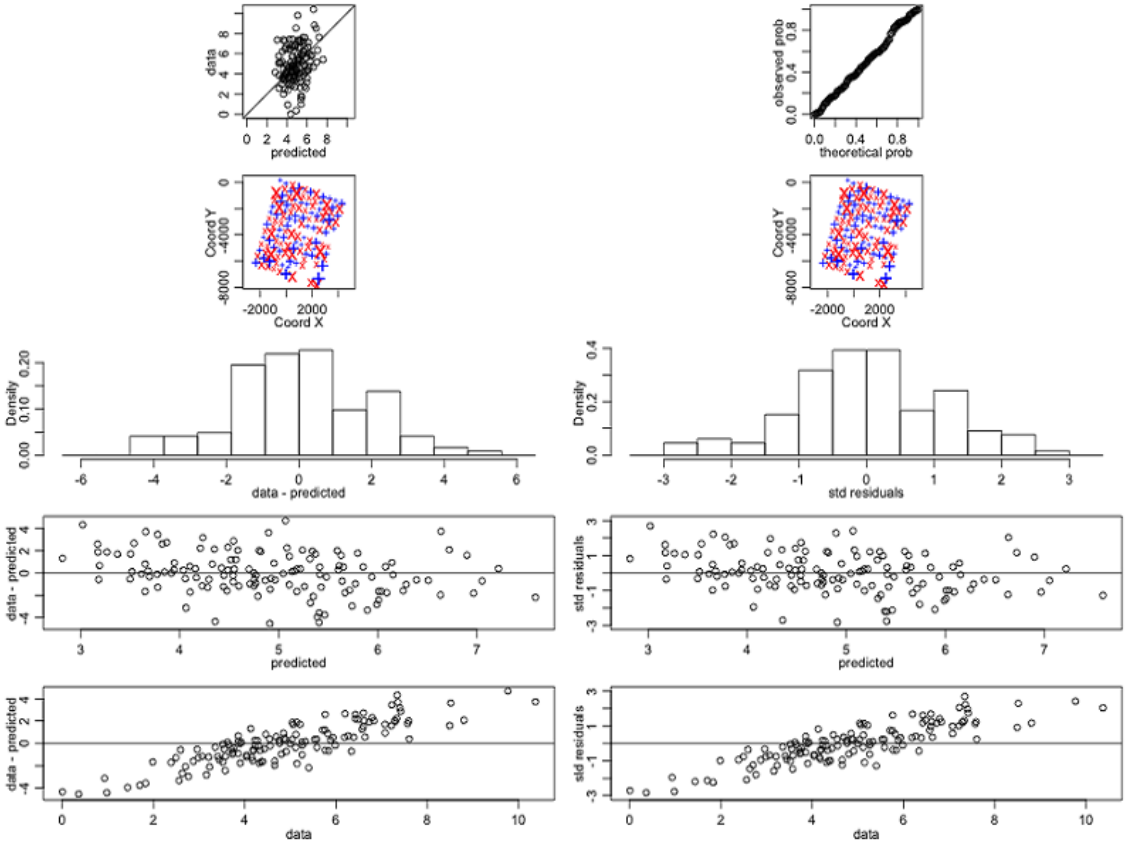
Fonte: Autor (2012)

Constata-se que os mapas gerados pelos métodos ML e REML com estrutura de covariância exponencial apresentaram superfícies muito semelhantes. Os modelos esféricos ML e REML também apresentaram superfícies bastante parecidas quando comparados. Os modelos 53 e 58 conseguem captar com mais detalhes a amplitude dos dados do ETENO.O modelo 53, estimado por ML com estrutura de covariância esférica, apresentou predições bem definidas e variâncias associadas às predições menores quando comparadas aos demais modelos. Portanto, acredita-se que o modelo 53 é o que melhor representa a estrutura espacial dos dados do ETENO.

4.3.5 Validação cruzada

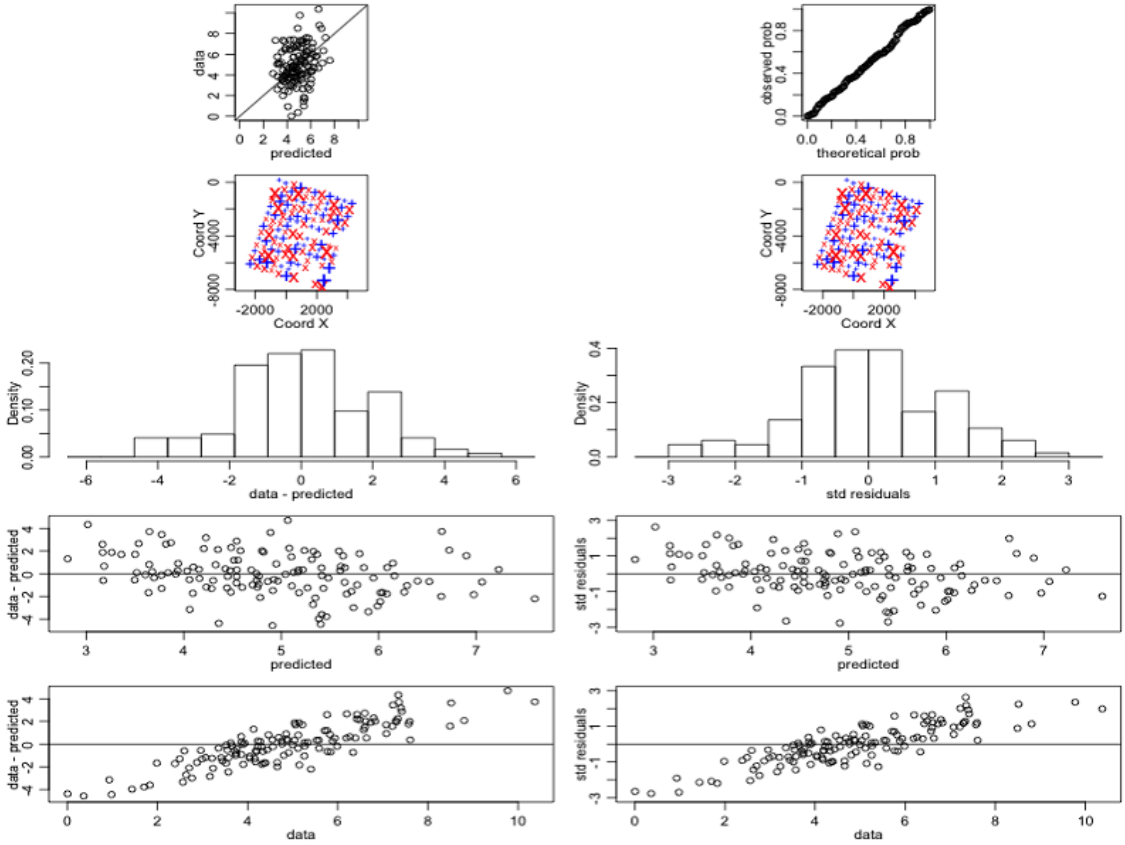
A validação cruzada para a variável ETENO apresentou resultados satisfatórios para todos os modelos selecionados. Verificou-se que os valores preditos estão próximo a reta padrão; a distribuição dos erros positivos e negativos está dispersa pela região de estudoe os dados da probabilidade teórica e observada se encontram sobre a reta, indicando que a predição foi eficiente através do uso do método da krigagem ordinária (Figura 45, Figura 46, Figura 47 e Figura 48).

Figura 45 - Validação cruzada para o ETENO referente ao modelo 43.



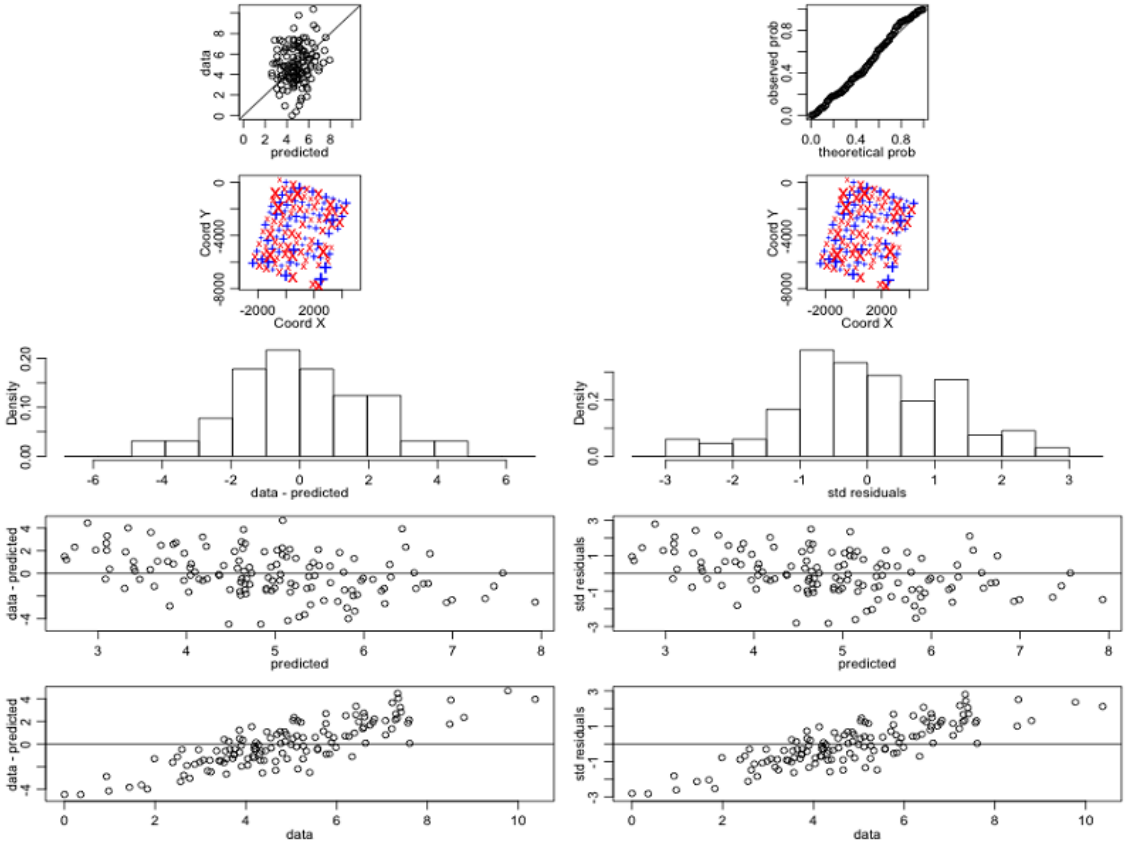
Fonte: Autor (2012)

Figura 46 - Validação cruzada para o ETENO referente ao modelo 48.



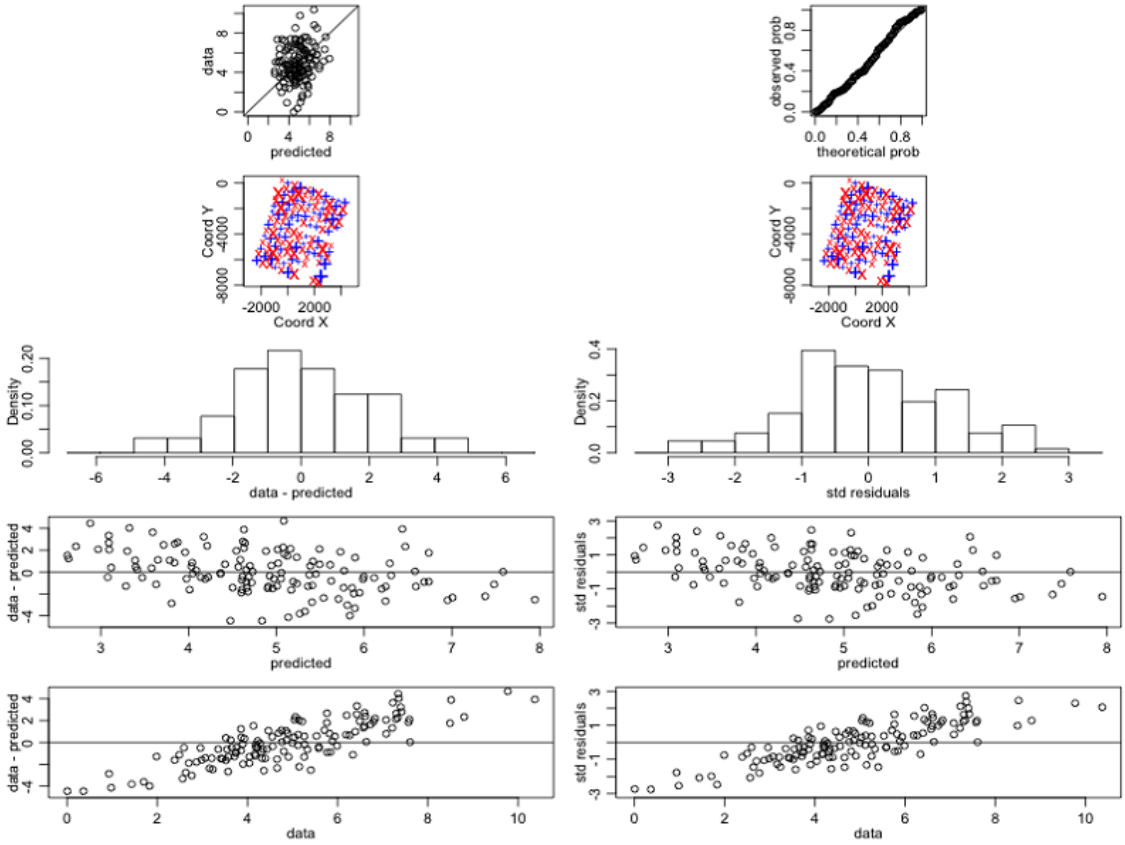
Fonte: Autor (2012)

Figura 47 - Validação cruzada para o ETENO referente ao modelo 53.



Fonte: Autor (2012)

Figura 48 - Validação cruzada para o ETENO referente ao modelo 58.



Fonte: Autor (2012)

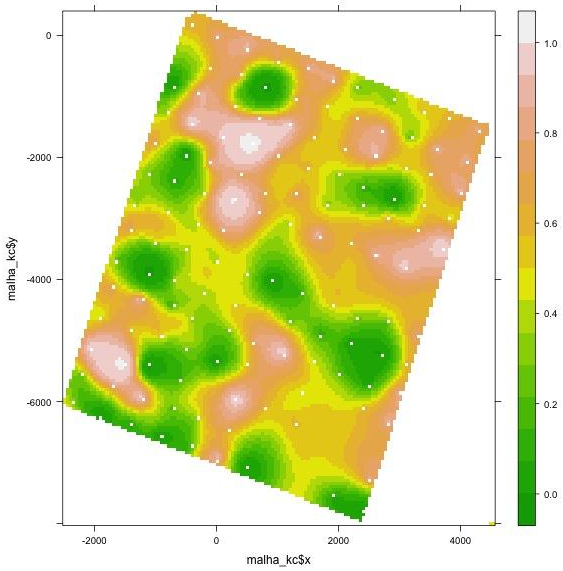
4.4 MAPAS DE PROBABILIDADES

Este trabalho tem como objetivo específico finalizar a análise geoestatística produzindo mapas de probabilidades envolvendo as distribuições espaciais dos gases analisados.

Na análise espacial realizada obtiveram-se, através da krigagem, os resultados das predições das estimativas pontuais para cada local da malha em estudo, juntamente com suas correspondentes estimativas de variância. Em posse destes valores, pode-se criar um vetor de dados, associados à sua respectiva localização com valores dos resíduos “studentizados” centrados por um valor específico, de interesse do pesquisador, que será denominado como ponto de corte. Desta maneira, como o processo é assumido como Normal, consegue-se associar probabilidades a cada ponto da malha de predição.

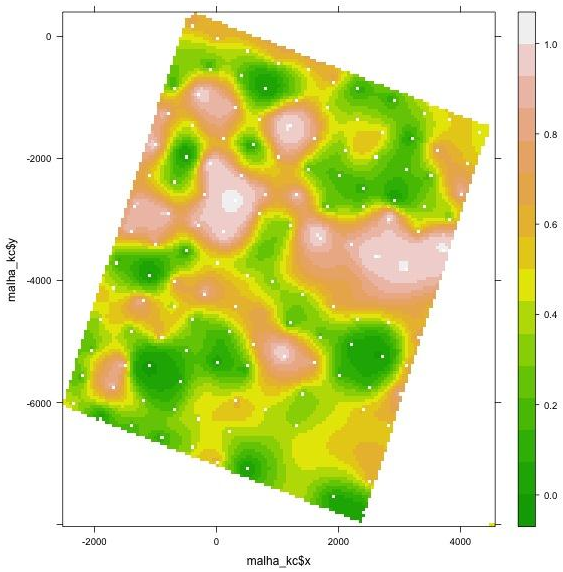
A Figura 49, Figura 50 e Figura 51 apresentam os mapas de probabilidades para as variáveis LOG_METANO, LOG_ETANO e ETENO, respectivamente, usando como ponto de corte a média estimada do processo. A escala de cores é apresentada à direita do gráfico e refere-se ao intervalo de probabilidade de 0 a 1, onde os valores em verdes estão associados a baixas probabilidades e as cores com tons mais claros estão associadas a altas probabilidades.

Figura 49 - Mapa de probabilidades do LOG_METANO resultado pelo modelo 18 com ponto de corte igual à média estimada do processo.



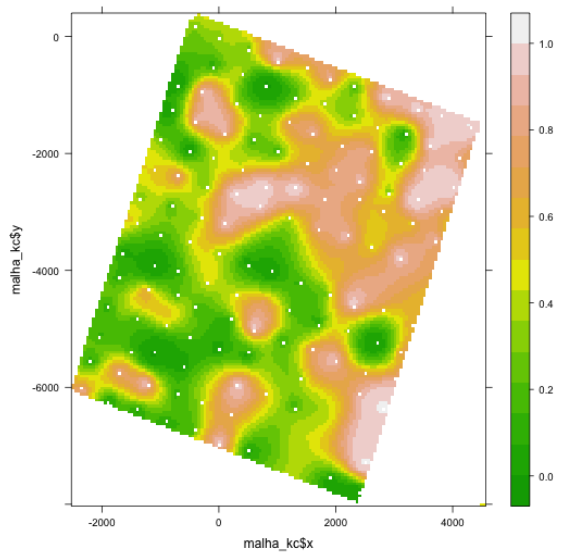
Fonte: Autor (2012)

Figura 50 - Mapa de probabilidades do LOG_ETANO resultado pelo modelo 38 com ponto de corte igual à média estimada do processo.



Fonte: Autor (2012)

Figura 51 - Mapa de probabilidades do ETENO resultado pelo modelo 53 com ponto de corte igual à média estimada do processo.

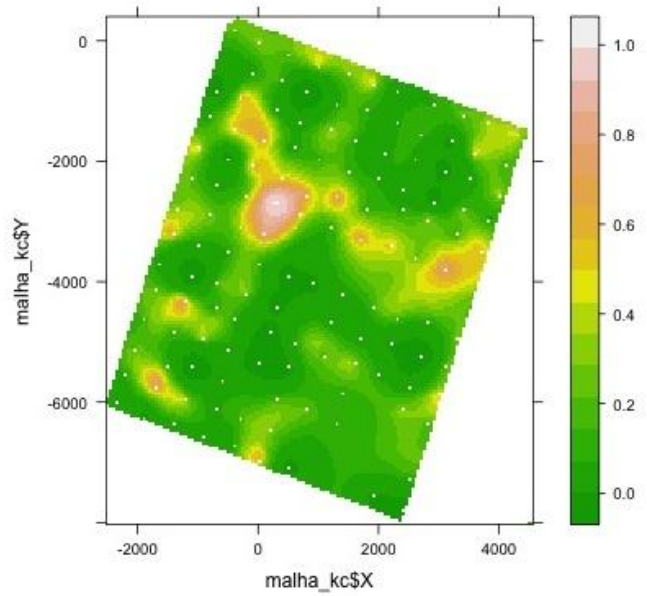


Fonte: Autor (2012)

O teorema do produto de probabilidades define que a densidade conjunta de eventos independentes é o produto de suas densidades marginais. Considerando independentes as distribuições espaciais de probabilidades dos gases analisados, é possível então criar mapas de probabilidades conjuntas para avaliar cenários de interesse.

A Figura 52 apresenta o gráfico com o mapa de probabilidades conjuntas utilizando a média do processo como ponto de corte. As densidades estimadas para cada variável referem-se aos modelos 18, 38 e 53, os quais foram os selecionados por melhor representarem os dados espaciais do LOG_METANO, LOG_ETANO e ETENO, respectivamente.

Figura 52 - Mapa de probabilidades conjuntas, estimados pelos modelos 18, 38 e 53, correspondendo às variáveis LOG_METANO, LOG_ETANO e ETENO, respectivamente. Média do processo utilizada como ponto de corte.



Fonte: Autor (2012)

Diferentes pontos de cortes podem ser utilizados para produção dos mapas de probabilidade. O pesquisador da área tem a flexibilidade de poder escolher, separadamente para cada variável modelada, qual valor adotar. Isso possibilita representar espacialmente diferentes cenários, favorecendo a identificação de regiões potencialmente favoráveis ou de interesse.

5 CONCLUSÃO

Este trabalho teve o objetivo de avaliar o comportamento da gasometria de superfície dos gases Metano, Etano e Eteno através do uso de métodos geoestatísticos em conjunto com a estimação por máxima verossimilhança, controlando importantes fatores físicos associados a fontes biogênicas.

Os métodos de estimação por máxima verossimilhança permitem a utilização de testes estatísticos para verificar a associação de covariáveis na distribuição dos dados de interesse quando adicionadas ao modelo. Constatou-se que a covariável Tipo de Solo apresentou associação e contribuiu significativamente, nos três modelos escolhidos, para representar a estrutura espacial dos gases metano, etano e eteno. Foi verificado também a tendência linear nas coordenadas na distribuição dos dados do eteno, a qual foi ajustada no referente modelo.

Os mapas de probabilidade trazem uma contribuição muito grande à análise ao possibilitarem que o fenômeno das exsudações dos gases possa ser analisado de forma integrada através da distribuição conjunta dos mesmos, assumindo que os processos são independentes. Esta técnica permite que pesquisadores façam uso de informações sobre as características de exsudações provenientes de fontes termogênicas relacionadas à gasometria de superfície, criando cenários que favoreçam a tomada de decisão sobre novas investidas exploratórias na região.

Portanto, identificou-se que ao inserir a covariável Tipo de Solo ao ajuste do modelo, foi possível gerar predições mais precisas sobre a distribuição espacial dos gases metano, etano e eteno na região em estudo. Consequentemente, permitiu-se gerar mapas de probabilidades também mais precisos, o que contribui para a diminuição do risco exploratório envolvido na prospecção de petróleo e gás.

Nos dados desta pesquisa, verificou-se alguma instabilidade computacional no processo de estimação por máxima verossimilhança quando utilizada a função *likfit()* do pacote *geoR*. Alguns critérios precisaram ser adotados para contornar esta instabilidade no momento em que resultados não satisfatórios eram obtidos. O auxílio do pesquisador, que tem conhecimento sobre o fenômeno, é muito importante para tratar destas escolhas que se fazem necessárias durante a modelagem, uma vez que foi necessária à intervenção sobre os valores estimados. Um destes critérios adotados foi o de delimitar o intervalo dos resultados das estimativas do parâmetro *phi*, algo que pode gerar alguma desconfiança e

desconforto por influenciar diretamente na estimativa da distância que dois pontos estão correlacionados. Em contra partida, ao conseguir modelar utilizando métodos de máxima verossimilhança, foi possível testar estatisticamente os fatores que podem influenciar na distribuição espacial dos dados, como foi de fato verificado.

Como sugestões de aprimoramentos e futuros estudos fica a proposta de: analisar outros hidrocarbonetos envolvidos na gasometria de superfície; realizar um estudo mais detalhado dos impactos gerados ao delimitar os intervalos de possíveis resultados para os parâmetros phi e sigmasq; validar a metodologia desenvolvida neste estudo acompanhado por um especialista na área de E&P.

REFERÊNCIAS

ALMEIDA, C. F. P.; RIBEIRO JR., P. J. (1996). *Estimativa da distribuição espacial de retenção da água em solo utilizando Krigagem Indicatriz.*Curitiba.

BOZDANGAN. H. (1987). *Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions.* Psychometrika. v.52, n.3.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data, Revised Edition.* New York: Wiley.

DIGGLE, P. J., RIBEIRO JR, P. J.; CHRISTENSEN, O. L. (2003). *An Introduction to Model-based Geoestatistics.* Spatial statistics and computational methods. Springer.

DIGGLE, P. J.; RIBEIRO JR., P. J. (2007). *Model-based Geostatistics.* New York: Springer.

_____. (2000). *Model based geostatistics.* Caxambu: ABE.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G. et al. (2004). *Análise Espacial de Dados Geográficos.* Brasília: EMBRAPA.

ISAAK, E. H.; SRIVASTAVA, R. M. (1989). *Applied geoestatistic: an introduction.* New York: Oxford University.

JOURNEL, A. G.; HUIJBREGTS, C. J. (1978). *Minning Geoestatistics.* London: Academic Press.

MATHERON, G. (1963). *Principles of Geoestatistics.* Economic Geology.

McBRATNEY, A., & WEBSTER, A. (1986). McBRATNEY, A.G.; WEBSTER, A.G. Choosing func- tions for semi-variograms and fitting them to sampling estimates. *Journal of Soil Science*, v. 37, p. 617-39.

MELLO, J. M.; BATISTA, J. L. F.; RIBEIRO JUNIOR, P. J.; OLIVEIRA, M. S. (2005). *Ajuste e seleção de modelos espaciais de semivariograma visand à estimativa volumétrica de Eucaliptus grandis. Scientia Forestalis.* Piracicaba.

MIRANDA, H. A. S. R. (2009). *Métodos robustos em geoestatística.*

ODA-SOUZA, M. M. (2009). *Modelagem geoestatística em quatro formações florestais do Estado de São Paulo.* Piracicaba.

PINHEIRO, J. B. (2000). Statistics and computing: Mixed-effects models in S and S-PLUS. New York: Springer.

SOARES, A. (2000). *Geoestatística para as ciências da terra e do ambiente, IST Press, Lisboa.* Lisboa: IST Press.

ANEXO A – Rotinas do Pacote geoR Utilizada nas Análises Realizadas neste Estudo

Código:

```
###################### Lendo banco de dados como objeto geodata
geoDB<- as.geodata(banco, head=T, coords.col=2:3, data.col=16 , covar.col=22:36)

#####################################################
########################### Modelo - Exponencial
patamar_parcial=0.15
alcance=1000
efeito_pepita=0.015
valor_lambda=1
modelo_variog="exponential"
metodo_estim="ML"
tendencia="cte"
lim_sigmasq_inf=0.01
lim_sigmasq_sup=0.5
lim_phi_inf=900
lim_phi_sup=1100

#####################################################
########################### Modelo - Esferico
patamar_parcial=0.15
alcance=1000
efeito_pepita=0.015
valor_lambda=1
modelo_variog="spherical"
metodo_estim="REML"
lim_sigmasq_inf=0.01
lim_sigmasq_sup=0.5
tendencia=as.formula(~Tipo_de_Solo_1+Tipo_de_Solo_2)

#####################################################
## Selecionar tendencia
tendencia=as.formula(~Umidade_1+Umidade_2)
tendencia=as.formula(~Tipo_de_Solo_1+Tipo_de_Solo_2)
tendencia=as.formula(~Cor_1+Cor_2+Cor_3)
tendencia=as.formula(~Uso_do_solo_1+Uso_do_solo_2+Uso_do_solo_3+Uso_do_s
olo_4)

## Fixa SILL
reml1= likfit(geoDB,trend=tendencia,fix.nug = fixar_nug ,cov.model =
modelo_variog,ini=c(patamar_parcial,alcance),nug=efeito_pepita,lik.method=metodo
_estim,limits = pars.limits(sigmasq=c(lim_sigmasq_inf, lim_sigmasq_sup)))
## Fixa PHI
```

```
reml1= likfit(geoDB,trend=tendencia,fix.nug = fixar_nug ,cov.model =
modelo_variog,ini=c(patamar_parcial,alcance),nug=efeito_pepita,lik.method=metodo
_estim,limits = pars.limits(phi=c(lim_phi_inf, lim_phi_sup)))
## Nug FIX
reml1= likfit(geoDB,trend=tendencia,fix.nug = fixar_nug ,cov.model =
modelo_variog,ini=c(patamar_parcial,alcance),nug=efeito_pepita,lik.method=metodo
_estim)
## Nug Livre
reml1= likfit(geoDB,trend=tendencia,cov.model =
modelo_variog,ini=c(patamar_parcial,alcance),lik.method=metodo_estim)

reml1
summary(reml1)
xv_reml1 <- xvalid(geoDB, model=reml1)
par(mfcol = c(5,2), mar=c(3,3,.5,.5), mgp=c(1.5,.7,0))
plot(xv_reml1, main="Cross validation")


#### Definindo tendencias a serem testadas ##########
tendencias<-cbind(tendencia=c("cte",
                        "~Umidade_1+Umidade_2",
                        "~Tipo_de_Solo_1+Tipo_de_Solo_2",
                        "~Cor_1+Cor_2+Cor_3",

"~Uso_do_solo_1+Uso_do_solo_2+Uso_do_solo_3+Uso_do_solo_4"))
tendencias
############################################################


#######Resumo Geral da Modelagem para as Tendencias Especificadas
############
n_sig=0.05
parametros_funcao<-
cbind(patamar_parcial,alcance,efeito_pepita,valor_lambda,modelo_variog,metodo_e
stim,tendencia,lim_sigmasq_inf,lim_sigmasq_sup,lim_phi_inf,lim_phi_sup,fixar_nug,fi
xar_ang,angulo)
colnames(parametros_funcao)<-
c("patamar_parcial","alcance","efeito_pepita","valor_lambda","modelo_variog","meto
do_estim","tendencia","lim_sigmasq_inf","lim_sigmasq_sup","lim_phi_inf","lim_phi_su
p","fixar_nug","fixar_ang","angulo")
resumo_geral=NULL
resumo_geral_betas=NULL
for( i in 1:length(tendencias)){
tendencia=tendencias[i]
if(tendencia!="cte") {
if(tendencia!="1st") {
if(tendencia!="2nd") tendencia=as.formula(tendencias[i])
}}
modeloML = likfit(geoDB,
trend=tendencia,
```

```
fix.nug = T ,
#                 psiA=1.3,
#                 fix.psiA=T,
#                 lambda=valor_lambda,
cov.model = modelo_variog,
ini=c(patamar_parcial,alcance),
nug=efeito_pepita,
lik.method=metodo_estim,
limits = pars.limits(sigmasq=c(lim_sigmasq_inf, lim_sigmasq_sup)),
#                 limits = pars.limits(phi=c(lim_phi_inf, lim_phi_sup)),
message=F)
resumo_geral<-parametros_estimados(modelo=modeloML,resumo=resumo_geral)
resumo_geral_betas<-
calcula_ic_betas(modelo=modeloML,n.sig=n_sig,nome.modelo=paste("MODELO",i),
resumo_betas=resumo_geral_betas,ini=2)
}
parametros_funcao
resumo_geral
resumo_geral_betas
#####################################################################


#### Exportar resumo ML Exponencial
#############################################################
write.table(t(resumo_geral), file="METANO_ML_EXP_resumo.csv",
row.names=T,sep=";",eol="\r\n")
write.table(resumo_geral_betas, file="METANO_ML_EXP_resumo_betas.csv",
row.names=T,sep=";",eol="\r\n")
#### Exportar resumo REML Exponencial
#############################################################
write.table(t(resumo_geral), file="METANO_REML_EXP_resumo.csv",
row.names=T,sep=";",eol="\r\n")
write.table(resumo_geral_betas, file="METANO_REML_EXP_resumo_betas.csv",
row.names=T,sep=";",eol="\r\n")
#### Exportar resumo ML Esferico
#############################################################
write.table(t(resumo_geral), file="METANO_ML_SPH_resumo.csv",
row.names=T,sep=";",eol="\r\n")
write.table(resumo_geral_betas, file="METANO_ML_SPH_resumo_betas.csv",
row.names=T,sep=";",eol="\r\n")
#### Exportar resumo REML Esferico
#############################################################
write.table(t(resumo_geral), file="METANO_REML_SPH_resumo.csv",
row.names=T,sep=";",eol="\r\n")
write.table(resumo_geral_betas, file="METANO_REML_SPH_resumo_betas.csv",
row.names=T,sep=";",eol="\r\n")

##### Cria Tabela com os parametros estimados de cada modelo
parametros_estimados<-function(modelo,resumo){
resumo<- cbind(Metodo=c(resumo[,1],modelo$method.lik),
```

```
                Modelo=c(resumo[,2],modelo$cov.model),
tendencia=c(resumo[,3],modelo$trend),
npars = c(resumo[,4],modelo$npars),
logL=c(resumo[,5],modelo$loglik),
                AIC=c(resumo[,6],modelo$AIC),
                BIC=c(resumo[,7],modelo$BIC),
nugget=c(resumo[,8],modelo$nugget),
sill=c(resumo[,9],modelo$cov.pars[1]),
range=c(resumo[,10],modelo$cov.pars[2]))
return(resumo)}
#####################################################################

###################### FUNCAO PARA CALCULO DOS IC DOS BETAS
###################
calcula_ic_betas<- function(modelo,n.sig,nome.modelo,resumo_betas,ini){
for(j in ini:length(modelo$beta)) {
resumo_betas<- cbind(Modelo=c(resumo_betas[,1],nome.modelo),
Beta_nome=c(resumo_betas[,2],ifelse(length(modelo$beta) ==
1,"Intercept",names(modelo$beta[j]))),
Beta_val=c(resumo_betas[,3],modelo$beta[j]),
Beta_IC_inf=c(resumo_betas[,4],modelo$beta[j] + qnorm(n.sig/2) *
ifelse(length(modelo$beta) == 1,sqrt(modelo$beta.var[j]),sqrt(modelo$beta.var[j,j]))),
Beta_IC_sup=c(resumo_betas[,5],modelo$beta[j] + qnorm(1-(n.sig/2)) *
ifelse(length(modelo$beta) == 1,sqrt(modelo$beta.var[j]),sqrt(modelo$beta.var[j,j]))
))
}
#rownames(resumo_betas) <-
ifelse(length(resumo_betas[,1])==1,paste("IC",n.sig*100,"%"),c(paste("IC",n.sig*100,"
%"),1:length(modelo$beta)))
return(resumo_betas)}
#####################################################################
```