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

Modelagem geoestatística da distribuição espacial da salinidade nas lavouras de arroz irrigado na porção leste do RS

Autor: Fernanda Rodrigues Vargas
Orientador: Professor Dr. Fernando Hepp Pulgati

Porto Alegre, 01 de Dezembro de 2010.

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

Modelagem geoestatística da distribuição espacial da salinidade nas lavouras de arroz irrigado na porção leste do RS

Autor: Fernanda Rodrigues Vargas

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Fernando Hepp Pulgati
Professora Dra. Jandyra Maria Guimarães Fachel

Porto Alegre, 14 de Dezembro de 2010.

AGRADECIMENTOS

Depois desses anos de estudos, quero dividir esse momento de alegria com as pessoas que estiveram ao meu lado e que fizeram parte desta conquista.

Agradeço a todos os meus professores pela dedicação e ensinamentos, em especial ao meu professor e orientador Fernando Pulgati pelas oportunidades e incentivo, à professora Jandyra Fachel pelo grande aprendizado que me proporcionou durante o trabalho no NAE, pelo incentivo e disponibilidade.

Agradeço à minha família pelo apoio e paciência, especialmente à minha mãe pelo esforço e educação que me dispensou, sempre com muita determinação. Ao Carlos pelo companheirismo e carinho, à Inára pelo incentivo aos estudos e dedicação. E a Clarissa Slongo pelo suporte e por me fazer acreditar que era possível.

Agradeço aos amigos que conquistei e que levarei comigo em meu coração e nas lembranças, pelos momentos de estudos em grupos, Silvana, Daniela, Marcel e Luciano (minha turma de Inferência II), e a todas as queridas colegas que dividiram grandes momentos comigo.

Agradeço a Deus por colocar todos vocês no meu caminho.

RESUMO

A Laguna dos Patos é a principal fonte de água para irrigação das lavouras nas planícies costeiras do Estado do RS. Devido à ligação da Laguna com o Oceano Atlântico essa água pode ocasionar o depósito de quantidades excessivas de sódio no solo, podendo prejudicar o estabelecimento da lavoura nos anos seguintes. No presente estudo, amostras de solo foram coletadas entre os meses de julho e setembro de 2008 e maio e agosto de 2009, a região de estudo compreendida entre os municípios de Rio Grande, localizado às margens da extremidade sul da Laguna dos Patos, e Torres, último município do Litoral Norte do Rio Grande do Sul, na divisa com Santa Catarina. A metodologia geoestatística foi utilizada para modelar a ocorrência de solos afetados pela salinidade pela análise das variáveis químicas representadas pelo teor de sódio trocável (Na), percentagem de sódio trocável (PST) e condutividade elétrica do extrato saturado (CE). A estrutura espacial foi explorada com o uso do semivariograma. Diferentes modelos foram ajustados através dos métodos de estimação de Mínimos Quadrados Ordinários e Ponderados, de Máxima Verossimilhança e Máxima Verossimilhança Restrita. A validação cruzada foi a técnica utilizada para comparar e verificar a qualidade da modelagem. As estimativas dos parâmetros do modelo selecionado foram usadas para predizer valores nos locais não amostrados através do método da Krigagem Ordinária. A influência da distância da margem da Laguna sobre os dados foi adicionada ao modelo como uma covariável, por isso a modelagem foi feita com e sem a covariável. O modelo Gaussiano foi escolhido para representar a dependência espacial dos dados. A significância estatística da covariável foi verificada para as variáveis Na e CE. Dentre os métodos utilizados, o de Máxima Verossimilhança Restrita apresentou o melhor resultado para as três variáveis.

Abstract

The Laguna dos Patos is the main source of water for irrigating crops in the coastal plains of RS state. Due to the connection of this Laguna with the Atlantic Ocean this water may contribute with excessive amounts of sodium deposition in soil that can hinder crop establishment in subsequent years. In the present study, soil samples were collected between July and September 2008 and May and August 2009, in an area extended between the municipalities of Rio Grande, located on the shores of the southern tip of Laguna dos Patos and Torres, last municipality of the North Coast of Rio Grande do Sul, on the border of Santa Catarina. The geostatistical methodology was used to model the occurrence of soils affected by salinity in the region through the analysis of the chemical variables represented by the content of exchangeable sodium (Na), exchangeable sodium percentage (ESP) and electrical conductivity of saturated extract (EC). The spatial structure was explored using the semivariogram. Different models were fitted using the estimation method of least squares Ordinary and the Weighted, Maximum Likelihood and Restricted Maximum Likelihood. Cross-validation technique was used to compare and verify the quality of modeling. The parameters of the selected model were used to predict values in areas not sampled by the method of Ordinary Kriging. The influence of distance from the edge of Laguna on the data was added to the model as a covariate, so the modeling was done with and without the covariate. The Gaussian model was chosen to represent the spatial dependence of data. Statistical significance of the covariate variables was observed for Na and CE. Among the different methods adopted, the Restricted Maximum Likelihood method showed the best result for all three variables.

SUMÁRIO

1. INTRODUÇÃO

As lavouras de arroz das planícies costeiras do Rio Grande do Sul podem apresentar problemas de salinidade no solo devido ao uso de água de má qualidade para a irrigação, água proveniente de rios litorâneos e de lagoas. A principal fonte de irrigação das lavouras nas planícies costeiras do Estado é a Laguna dos Patos. Sua ligação com o oceano Atlântico, no município de Rio Grande, acarreta o transporte de água salina para dentro da Laguna, o que afeta a característica da água usada na irrigação. A quantidade de água do mar ocorre com mais ou menos intensidade no decorrer do ano de acordo com as mudanças das estações. No inverno, por exemplo, ocorre um aumento na ocorrência mais frequente de ventos na região.

Compreender como ocorre a salinidade e como ela se distribui na região de estudo, lavouras de arroz da costa gaúcha, contribui para o uso mais eficiente e adequado do solo, podendo servir de base para ações de tratamento da salinidade quer para plantações futuras, quer para manutenção da qualidade do solo. Através da modelagem dos dados é possível mapear a ocorrência e a distribuição dos componentes químicos que caracterizam o solo afetado por sais.

A monografia foi construída a partir do trabalho dos pesquisadores Felipe Carmona[1], Ibanor Anghinoni[2] e Eliseu José Weber[3]. A definição do plano amostral e o procedimento de coleta dos dados foram realizados pelos autores da pesquisa, tratando o presente trabalho, portanto, da modelagem do problema. Os dados foram coletados entre os meses de julho e setembro de 2008 e maio e agosto de 2009. A região de estudo compreende a Planície Costeira Interna e Externa da Laguna dos Patos. A amostragem se estendeu entre os municípios de Rio Grande, localizado às margens da extremidade sul da Laguna dos Patos, e Torres, último município do Litoral Norte do Rio Grande do Sul, na divisa com Santa Catarina.

Foram observados 766 pontos nas áreas que são, ou podem vir a ser, usadas para o cultivo de arroz irrigado por inundação.

[1] Eng. Agrônomo, Doutorando do Programa de Pós Graduação em Ciência do Solo, UFRGS. Trabalho em desenvolvimento: "SOLOS AFETADOS POR SAIS NAS PLANÍCIES COSTEIRAS DO RIO GRANDE DO SUL".
[2] Professor Adjunto, Programa de Pós Graduação em Ciência do Solo, UFRGS.
[3] Pesquisador, Laboratório de Geoprocessamento do Centro de Ecologia, UFRGS.

Assumir que os dados foram obtidos sob as mesmas condições e de maneira independente, formando uma amostra aleatória (dados independentes e identicamente distribuídos), torna a teoria matemática estatística mais tratável. No entanto, para os dados espaciais a dependência está presente em todas as direções, não podendo ser modelados como independentes estatisticamente. Observações próximas têm um comportamento mais similar, então na medida em que a distância entre essas observações aumenta essa similaridade tende a diminuir; dessa forma é preciso identificar a dependência espacial através de medidas de associação, como a autocorrelação espacial.

A Estatística Espacial é um ramo da ciência Estatística que quantifica as propriedades e o relacionamento de um fenômeno de interesse considerando sua localização espacial. Cressie (1993) a divide em três partes de acordo com o tipo de dado: processos espaciais indexados sobre o espaço contínuo (geoestatística), processos espaciais indexados sobre lattices no espaço (dados de área, espaço análogo de séries temporais) e processos espaciais pontuais.

Os dados deste trabalho serão analisados através da estatística espacial sob enfoque da geoestatística, visto que na superfície, ou região de estudo, a salinidade se distribui continuamente no espaço.

O objetivo da modelagem geoestatística é a estimação e predição, sendo o primeiro relacionado à inferência dos parâmetros de um modelo que expresse a dependência espacial, e o outro, em predizer valores não observados na região de estudo com base nesse modelo. Alguns exemplos deste tipo de dado são: concentração de poluentes, medidas de chuva, propriedades do solo, entre outros.

O engenheiro de minas Daniel Krige (1951) ao realizar estudos com dados de concentração de ouro de minas do Rand, na África do Sul, concluiu que apenas a informação dada pela variância não era suficiente para explicar o fenômeno estudado, sendo necessária incorporar a informação da distância entre as observações. Naquela época a aplicação da geoestatística era voltada para a estimação das reservas de minério.

O desenvolvimento teórico da geoestatística ocorreu dos estudos de George Matheron e de seus colegas da escola francesa - *Centre du Morphologie Mathématique* - em Fontainebleau, França, baseado nos resultados de Krige. O conceito da *Teoria das Variáveis Regionalizadas*, Matheron (1963), foi então fundamentado, dando a importância adequada para as relações

espaciais existentes entre as observações. A aplicação dessa teoria a problemas na geologia e mineração conduziram ao nome Geoestatística.

A partir da década de 70 sua utilização passou a ser mais abrangente, ao mesmo tempo em que seus conceitos evoluíram através de outros autores como Cressie (1993).

2. METODOLOGIA GEOESTATÍSTICA

A geoestatística está interessada no modo como os dados variam no espaço contínuo dentro da área de estudo. As localizações espaciais são fixas podendo estar distribuídas regularmente ou irregularmente. Um dos objetivos sob este enfoque é predizer valores em locais não amostrados, possibilitando gerar uma superfície contínua que expresse a distribuição do fenômeno estudado sobre toda a região, no caso desta monografia o fenômeno de interesse é a salinidade.

Para fazer a análise dos dados é necessário definir o modelo geoestatístico. A seguir os conceitos da geoestatística são apresentados.

O foco é estudar o fenômeno espacial $w(s)$ que existe sobre toda a região em estudo $\mathcal{D}$, $\mathcal{D} \subset \Re^2$, e $w(s)$ é tratado como uma realização de um processo estocástico $W(\cdot) = \{W(s): s \in \mathcal{D}\}$. Porém $W(\cdot)$ não é observável diretamente. Em vez disso, os dados consistem das medidas $z_1, \dots, z_n$, relacionadas à determinada variável de interesse Z, tomadas em suas respectivas localizações $s_1, \dots, s_n$ amostradas dentro da região $\mathcal{D}$, e Z_i é uma versão ruído de $W(s_i)$. O delineamento amostral para $s_1, \dots, s_n$ pode ser assumido ser determinístico ou estocasticamente independente do processo $W(\cdot)$ que gera z_i, e todas as análises realizadas são condicionadas às localizações $s_1, \dots, s_n$. (Diggle, Ribeiro e Christensen, 2003)

Os dados geoestatísticos têm a seguinte forma:

$$(s_i, z_i): i = 1, \dots, n \qquad (2.1)$$

onde s_i identifica a localização espacial (tipicamente em duas dimensões) dentro da região de estudo $\mathcal{D} \subset \Re^2$, e z_i é o valor da variável associada à localização s_i.

Assumimos que z_i é uma realização da variável aleatória Z_i cuja distribuição é dependente do valor de s_i de um processo estocástico contínuo subjacente $W(s)$ não observável diretamente. A quantidade observável Z_i é diferente do processo latente $W(s)$. O valor de z_i ocorre em qualquer ponto no espaço dentro da região $\mathcal{D}$ e pode ser generalizada por inferência, na forma espacialmente contínua, para toda a região observada.

O modelo geoestatístico, segundo Diggle & Ribeiro (2000), é definido através de um processo estocástico $Z(s): s \in \mathcal{D}$, que é uma realização parcial de um processo estocástico da

forma $Z(s)$: $s \in \Re^2$. Os valores de Z_i podem ser considerados uma versão ruído do processo estocástico subjacente $W(s_i)$, *signal process*, e Z_i pode ser assumido condicionalmente independente de $W(\cdot)$. Assim, a forma básica de um modelo geoestatístico incorpora no mínimo duas componentes: um processo estocástico $Z(s)$ e um modelo estatístico para as variáveis aleatórias $Z = (Z_1, \dots, Z_n)$ condicionadas a $\{W(s): s \in \mathcal{D} \subset \Re^2\}$.

Resumindo: Seja $Z(s)$ a variável aleatória medida na localização $s \in \mathcal{D} \subset \Re^2$, onde $\mathcal{D}$ é a região de estudo e *z*(*s*) é uma realização da variável. O processo estocástico é definido como uma família de variáveis aleatórias reais indexadas pela posição e que variam continuamente no espaço.

O modelo geoestatístico apresenta limitações que requerem suposições que possibilitem fazer inferências estatísticas seguras. Uma destas limitações se deve ao fato de que os dados geoestatísticos representam apenas uma realização do processo estocástico, ou seja, n amostras de tamanho um identificadas em suas localizações, assim necessitaríamos de várias realizações desta variável aleatória para gerar uma amostra suficientemente grande para inferir a respeito do fenômeno.

Um campo aleatório é um processo estocástico que existe em algum espaço de dimensão d, sua definição é dada por $\{Z(s_i): s_i \in \mathcal{D} \subset \Re^2\}$, sendo $Z(s_i)$ a variável aleatória que varia continuamente em $\mathcal{D}$, onde s é a localização da variável, $\mathcal{D}$ é a região de estudo e $\Re^d$ é o espaço de dimensão d. A geoestatística modela os valores z_i dentro de $\mathcal{D}$ como um processo estocástico $\{Z(s_i): s_i \in \mathcal{D} \subset \Re^2\}$. Em cada posição espacial tem-se um valor de z_i que é uma variável aleatória e cada z_i assume uma distribuição de probabilidade.

A amostra coletada representa um número potencialmente infinito de medidas $\{z(s): s \in \mathcal{D}\}$ que podem ser tomados por toda região de estudo $\mathcal{D}$. O desenvolvimento da modelagem é baseado na suposição de que $\{z(s): s \in \mathcal{D}\}$ é uma realização de um processo estocástico

$$\{Z(s): s \in \mathcal{D}\}, \qquad \mathcal{D} \in \Re^d \tag{2.2}$$

onde $\mathcal{D}$ é um subconjunto fixo de $\Re^d$ com $d = 1,2, \dots$ (Cressie, 1993).

Para enfatizar a fonte de aleatoriedade, o processo estocástico (2.2) é escrito às vezes como $\{Z(s; \nu): s \in \mathcal{D}; \nu \in \Omega\}$ onde $(\Omega, \mathcal{F}, P)$ é o espaço de probabilidade. A realização de $\{z(s): s \in \mathcal{D}\}$ deveria corresponder a um particular valor de $\nu = \nu_0$.

Matheron (1963) chama a quantidade $z(\cdot)$ de variável regionalizada para enfatizar o aspecto contínuo do fenômeno. Toda variável distribuída no espaço e que apresenta uma estrutura de correlação espacial é regionalizada. Como $z(s)$ é o valor da variável aleatória Z na localização s, então $Z(s)$ é uma variável regionalizada.

O processo aleatório (2.2) é geralmente definido através da distribuição de dimensão finita da forma

$$F_{s_1,\dots,s_m}(z_1,\dots,z_m) \equiv P\{Z(s_1) \le z_1,\dots,Z(s_m) \le z_m\}, \qquad m \ge 1 \tag{2.3}$$

que deve satisfazer as condições de simetria e consistência de Kolmogorov, Cressie (1993).

Suponha $\mu(s) \equiv E(Z(s))$ existe para todo s $\in \mathcal{D}$; chamamos $\mu(\cdot)$ de tendência (ou *drift*). A existência de var$(Z(s))$ para todo s $\in \mathcal{D}$ permite definir estacionariedade de segunda-ordem e estacionariedade intrínseca.

Estacionariedade de Segunda-Ordem

Algumas suposições sobre Z devem ser feitas, caso contrário os dados representariam uma amostragem incompleta de uma simples realização, sendo impossível fazer inferências. Pode-se assumir que

$$E\big(Z(s)\big) = \mu, \qquad \text{para todo } s \in \mathcal{D} \tag{2.4}$$

A média do processo é constante para todo espaço contínuo $\mathcal{D}$ independente da variação da localização espacial s. Ou que $F_s(z) \equiv \Pr(Z(s) \le z)$ não depende de s. A fim de estimar os preditores lineares ótimos, adiciona-se à suposição

$$\text{cov}\left(Z(s_1), Z(s_2)\right) = C(s_1 - s_2), \qquad \text{para todo } s_1, s_2 \in \mathcal{D} \tag{2.5}$$

A função $C(\cdot)$ é chamada de covariograma ou uma função covariância estacionária.

Definição: Uma função aleatória $Z(\cdot)$ satisfazendo as condições (2.4) e (2.5) é definida ser *estacionária de segunda-ordem*, ou *estacionária fraca*. Além disso, se $C(s_1 - s_2)$ é somente uma função da distância existente entre dois locais, $\| s_1 - s_2 \|$, é dito *isotrópico*.

Estacionariedade Intrínseca

Uma alternativa menos restritiva, quando a estacionariedade de segunda-ordem não é verificada, é assumir a estacionariedade intrínseca que é definida pela diferença entre as variáveis aleatórias separadas pela distância h.

Supondo que a amostra em um ponto s, $s \in \Re^d$ $(d = 1,2, \dots)$, é uma realização de um processo estocástico (2.2) e que esta é observada em certos pontos $\{s_i : i = 1, \dots, n\}$ da região $\mathcal{D}$, então a estacionariedade intrínseca é definida por

$$E\big(Z(s+h) - Z(s)\big) = 0 \tag{2.6}$$

$$\text{var}\,\big(Z(s+h) - Z(s)\big) = 2\gamma(h) \tag{2.7}$$

onde $h = \| s_j - s_i \|$ é o vetor da diferença entre as localizações i e j, e a quantidade $2\gamma(h)$ é conhecida como *variograma*, parâmetro fundamental da geoestatística, sendo a principal ferramenta para todo o processo de estimação e predição dos dados.

A variabilidade entre duas observações depende somente da distância entre elas, isto é, $C(s_1 - s_2)$ é uma função somente da distância euclidiana $h = \| s_1 - s_2 \|$.

O estimador clássico do variograma proposto por Matheron (1963) é

$$2\hat{\gamma}(h) \equiv \frac{1}{|N(h)|}\sum_{N(h)}(Z(s_i) - Z(s_j))^2 \tag{2.8}$$

onde o somatório está definido sobre $N(h) \equiv \{(i,j) : s_i - s_j = h\}$ e $|N(h)|$ é o número de elementos diferentes de $N(h)$. Este estimador é não viesado, porém possui propriedades pobres; são afetados por observações atípicas, *outliers*, devido o termo ao quadrado em (2.8). (Cressie, 1993)

Cressie e Hawkins (1980) propuseram um estimador mais robusto para a estimação do variograma (mais detalhes em Cressie, 1993)

$$2\bar{\gamma}(\mathrm{h}) \equiv \frac{\left\{\frac{1}{|\mathrm{N(h)}|}\sum_{\mathrm{N(h)}}|\mathrm{Z(s_i)} - \mathrm{Z(s_j)}|^{\frac{1}{2}}\right\}^4}{\left(0.457 + \frac{0.494}{|\mathrm{N(h)}|}\right)} \tag{2.9}$$

Ergodicidade

Existe um subconjunto das funções aleatórias estacionárias de segunda-ordem que possui uma propriedade fundamental conhecida como *ergodicidade*. Ao assumir estacionariedade no processo $Z(s)$ os primeiros momentos podem ser estimados a partir desta propriedade. A função densidade de probabilidade do processo pode ser estimada a partir dos dados em toda a região $\mathcal{D}$, assumimos que o comportamento da função densidade de probabilidade para uma realização do processo estocástico é idêntico ao da função densidade de probabilidade global.

Permite uma estimação consistente a partir de amostras retiradas de uma única realização para o fenômeno contínuo gerado aleatoriamente. É assumida a suposição de que a média de uma única realização do processo dentro da região $\mathcal{D}$ é igual à média de todas as possíveis realizações da variável Z. Com isso a propriedade de *ergodicidade* nos permite utilizar os dados amostrais e com isso contorna os problemas relativos à falta de repetição que prejudica as inferências.

Uma abordagem mais detalhada sobre *ergodicidade* pode ser encontrada em Cressie (1993).

Isotropia

Supondo que um processo aleatório $Z(\cdot)$ satisfaz: $E\big(Z(s)\big) = \mu$, para todo $s \in \mathcal{D}$, e $\text{cov}\big(Z(s_1), Z(s_2)\big) = C(s_1 - s_2)$, para todo $s_1, s_2 \in \mathcal{D}$, então ele é dito estacionário de segunda-ordem, e nesse caso também é intrinsecamente estacionário. Se o covariograma $C(\cdot)$ depende apenas da distância euclidiana entre dois locais, $\| s_1 - s_2 \|$, o processo é chamado *isotrópico*.

No contexto aplicado, o fenômeno é dito *isotrópico* quando a variabilidade espacial, expressa pelo variograma, é a mesma em todas as direções.

Anisotropia

O processo é dito ser *anisotrópico* quando a dependência entre $Z(s)$ e $Z(s+h)$ é uma função da distância e direção de $h = \| s_1 - s_2 \|$, ou seja, o variograma não é apenas uma função da distância entre duas localizações espaciais. A variabilidade espacial do fenômeno não é a mesma em todas as direções.

Anisotropias são causadas por processos físicos subjacentes envolvendo diferenciabilidade no espaço, Cressie (1993). Algumas vezes a anisotropia pode ser corrigida pela transformação linear do vetor de *lag* h.

Processo Gaussiano

Processos estocásticos gaussianos são amplamente usados na prática como modelos para dados geoestatísticos. A suposição gaussiana é o modelo base (*model-based*) de alguns métodos de predição amplamente usados na geoestatística, como Krigagem Ordinária e Universal.

Quando o processo estocástico é gaussiano a estacionariedade de segunda-ordem e a estacionariedade forte coincidem, porque um processo gaussiano é caracterizado pela sua média e sua função de covariância. Uma condição suficiente para ergodicidade nesse caso é $C(h) \to 0$, quando $\| h \| \to \infty$ (Adler, 1981, p.145).

Processos gaussianos são importantes por duas razões, segundo Cressie (1993). A primeira razão é que sobre a suposição de um processo gaussiano os processos de estimação e predição empregando a teoria de distribuições são simplificados. A segunda diz respeito às considerações assintóticas onde efeitos de variação de pequena escala podem ser aproximados pela distribuição Gaussiana devido às propriedades do Teorema do Limite Central (Lingren, 1976, p.157).

Um processo estocástico é gaussiano para qualquer localização $s_i \in \Re^2$ $(i = 1, \dots, n)$ se a distribuição conjunta de $Z = Z(s_1), \dots, Z(s_n)$ é Gaussiana multivariada. Assim, a estacionariedade de segunda-ordem é assumida se a média e a variância de $Z(s)$ são constantes para todo s $\in \Re^2$, e a função de covariância entre duas localizações depende apenas da distância entre elas.

As suposições do modelo gaussiano subjacente são:

1. O processo estocástico $\{W(s): s \in \Re^2\}$ é um processo gaussiano com média μ, variância σ^2 =var $\{W(s)\}$ e função de correlação $\rho(h)$ =corr $\{W(s), W(s')\}$, onde $h = \| s - s' \|$ e $\|\cdot\|$ denota a distância euclidiana.

2. Condicionados a $\{W(s): s \in \Re^2\}$, z_i são realizações mutuamente independentes da variável aleatória Z_i, normalmente distribuídos com média condicional $E[Z_i|W(\cdot)] = W(s_i)$ e variância condicional τ^2.

O modelo pode ser definido como (Diggle & Ribeiro, 2007)

$$Z(s_i) = W(s_i) + Y_i, \quad i = 1, \dots, n \tag{2.10}$$

onde $\{W(s): s \in \Re^2\}$ é definido na suposição 1 acima, e Y_i são variáveis aleatórias mutuamente independentes com distribuição $N(0, \tau^2)$.

O modelo acima é válido se a função de correlação $\rho(h)$ for não-negativa, essa restrição assegura variância não-negativa. (Diggle & Ribeiro, 2007, p.29).

Modelagem da Estrutura de covariância espacial

A covariância na Estatística elementar mede a relação entre duas variáveis distintas, no contexto espacial essa medida é feita sobre a mesma variável em localizações diferentes. A análise da estrutura espacial ajuda a compreender como ocorre o fenômeno no espaço e como as observações se relacionam entre si considerando suas localizações espaciais. O variograma quantifica a dependência espacial entre duas variáveis regionalizadas e a correlação espacial entre elas, é uma alternativa para a caracterização de dependência de segunda-ordem em um processo estocástico.

Suponha

$$\text{var}\,[Z(s_1) - Z(s_2)] = 2\gamma(s_1 - s_2), \quad \text{para todo } s_1, s_2 \in \mathcal{D} \tag{2.11}$$

A função que depende somente dos incrementos $s_1 - s_2$, foi denominado por Matheron (1963) de variograma, e é expresso pela quantidade $2\gamma(\cdot)$. Assumindo que a função 2γ existe, Cressie (1993) a tratou como um parâmetro do processo aleatório $Z(\cdot)$. A restrição aplicada ao variograma é que deve ser não-negativo.

O variograma é uma ferramenta básica de suporte às técnicas de geoestatística, que permite representar quantitativamente a variação de um fenômeno regionalizado no espaço (Huijbregts, 1975).

Alguns autores utilizam o semivariograma $\gamma(\cdot)$, definido a seguir, para expressar a dependência espacial.

$$2\gamma(h) = \text{var}\left(Z(s_1) - Z(s_2)\right) = 2C(0) - 2C(h) \tag{2.12}$$

dividindo (2.12) por 2, temos o semivariograma que é análogo ao variograma

$$\gamma(h) = C(0) - C(h) \tag{2.13}$$

$$C(0) = \text{var}\,[Z(s_i)] = C(s_i - s_i) \tag{2.14}$$

Relembrando que o processo gaussiano é assumido ser intrinsecamente estacionário, o desenvolvimento algébrico está considerando os resultados apresentados.

Efeito Pepita

Supondo isotropia, $\gamma(h) = \gamma(-h)$ e $\gamma(0) = 0$. Se $\gamma(h) \to c_0 > 0$ quando $h \to 0$, então c_0 é chamado *efeito pepita* segundo Matheron (1963), ele acredita que uma variação em microescala esteja causando uma descontinuidade na origem do semivariograma. Matematicamente isto não pode acontecer já que o processo $Z(\cdot)$ é contínuo então $E[Z(s+h) - Z(s)]^2 \to 0$, quando $\| h \| \to 0$. A possível explicação para a descontinuidade $c_0 > 0$ é que esta seja uma medida de erro, chamada de c_{ME}. Como somente possuímos os dados e nada pode ser dito sobre o variograma no *lag* de distância menor do que $\min\{\| s_i - s_j \| : i \leq 1 < j \leq n\}$. Sob o enfoque matemático, para modelar o processo com variações de microescala é adicionado um "ruído branco" com média zero, variância constante e covariância zero. Isto é uma suposição verificada em amostras geograficamente próximas. Chamam a variância do "ruído branco" de c_{MS}, que representa o efeito pepita do processo de microescala. Então

$$c_0 = c_{MS} + c_{ME} \tag{2.15}$$

O comportamento do semivariograma na origem é muito informativo a respeito das propriedades do processo estocástico $Z(\cdot)$. Cressie (1993) cita os tipos mais comuns destes comportamentos baseado na categorização de Matheron (1963).

Definição: Suponha que $\{Z(s) : s \in \mathcal{D}\}$ satisfaça (2.4) e (2.11). Então $Z(\cdot)$ é dito ser intrinsecamente estacionário, ou dizendo de outra maneira, que satisfaz a hipótese intrínseca.

Além disso, se $2\gamma(s_1 - s_2)$ é uma função somente da distância euclidiana de dois pontos locais $\| s_1 - s_2 \|$, então $2\gamma(\cdot)$ é chamado de isotrópico.

Covariograma e correlograma

A função C(·) dada em (2.5) é chamada covariograma (ou função de autocovariância). Se $C(0) > 0$, dizemos que

$$\rho(h) \equiv \frac{C(h)}{C(0)} \tag{2.16}$$

é o correlograma (ou função de autocorrelação). Sabemos que $C(h) = C(-h)$, $\rho(h) = \rho(-h)$ e $\rho(0) = 1$. Considere a relação

$$\text{var}\left(Z(s_1) - Z(s_2)\right) = \text{var}\,[Z(s_1)] + \text{var}\,[Z(s_2)] - 2\,\text{cov}\,[Z(s_1), Z(s_2)] \tag{2.17}$$

$$\text{var}\left(Z(s_1) - Z(s_2)\right) = 2\gamma(s_1 - s_2) \tag{2.18}$$

Se $Z(\cdot)$ é um processo estacionário de segunda-ordem, a variância de duas variáveis aleatórias de localizações distintas é $\text{var}\left(Z(s_1) - Z(s_2)\right) = 2(C(0) - C(s_1 - s_2))$, que implica que $Z(\cdot)$ é intrinsecamente estacionário com

$$2\gamma(h) = 2(C(0) - C(h)) \tag{2.19}$$

$$\gamma(s_1 - s_2) = C(0) - C(s_1 - s_2) \tag{2.20}$$

como a variância do processo $Z(\cdot)$ é igual a $C(0)$ e $h = (s_1 - s_2)$

$$\gamma(h) = \sigma^2 - C(h) \tag{2.21}$$

Então, em um processo estacionário o semivariograma (ou variograma) é definido pelo inverso do covariograma (Figura1). O covariograma inicia em $C(0) = \sigma^2$, decrescendo para zero na medida em que a distância dada por h aumenta, enquanto o semivariograma inicia em zero ou em um ponto no eixo y (efeito pepita) e cresce continuamente com o aumento da distância até o valor igual a σ^2. O comportamento do semivariograma indica que quanto menor a distância entre as observações, mais similares elas são.

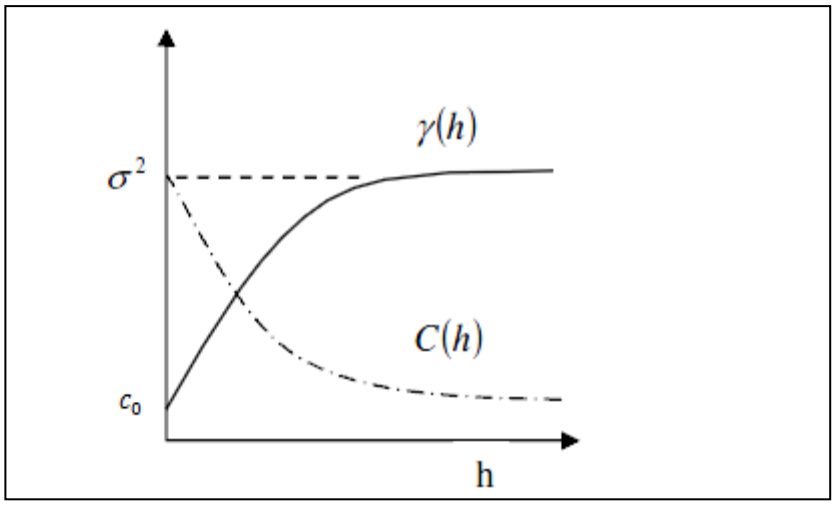

Figura 1 - Covariograma e Semivariograma

Se $C(0) \to 0$, quando $\| h \| \to \infty$, como por exemplo quando $Z(\cdot)$ é um processo gaussiano estacionário ergódico, então $\gamma(h) \to C(0)$. A quantidade $C(0)$ é chamada de patamar ou *sill* do semivariograma. O patamar parcial ou *partial sill* é definido como $C(0) - c_0$, onde c_0 é o efeito pepita ou *effect nugget* como definido em (2.15). O menor valor de $\| r_0 \|$ no qual $2\gamma(r_0(1+\epsilon)) = 2C(0)$, para qualquer $\epsilon > 0$, é chamado alcance ou *range* do semivariograma na direção $r_0/\| r_0 \|$.

A Figura 2 identifica os parâmetros do semivariograma citados acima em um gráfico onde no eixo das ordenadas estão os valores para o semivariograma gerados a partir dos dados amostrais pelo estimador (2.8) ou (2.9) na medida em que a distância h aumenta, e no eixo das abcissas estão identificadas as distâncias entre a diferença dos pares de valores amostrais.

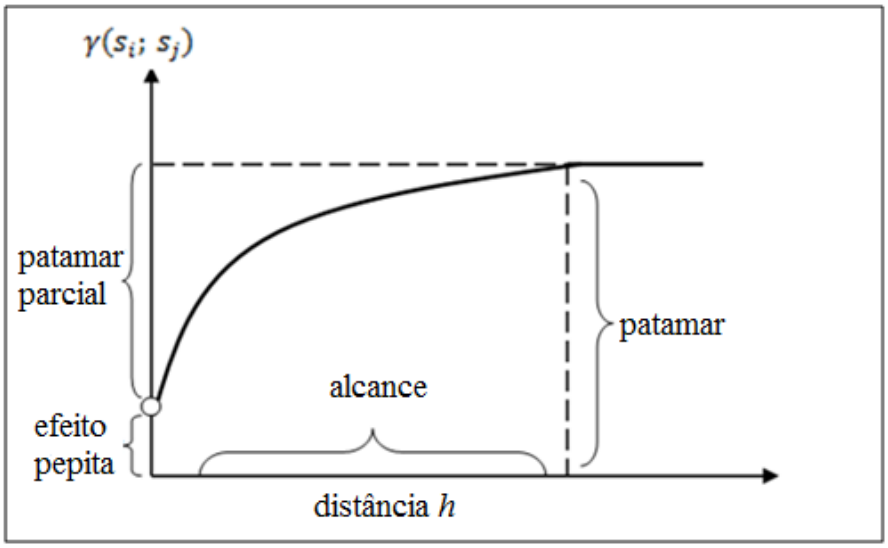

Figura 2 – Parâmetros do semivariograma.

A classe de todos os processos estacionários de segunda-ordem estão rigorosamente contidos na classe de todos os processos intrinsecamente estacionários, Cressie (1993, p.68).

Para o modelo linear gaussiano (2.10), com $h = \| s_1 - s_2 \|$, a variância é dada por

$$\text{var}\,(h) = \tau^2 + \sigma^2\{1 - \rho(h)\} \qquad (2.22)$$

Os parâmetros básicos da covariância do modelo linear gaussiano são a *nugget variance*, τ^2, o *total sill*, $\tau^2 + \sigma^2 = \text{var}\,[Z(s)]$, e o *range*, ϕ, que vem de $\rho(h) = \rho_0(h/\phi)$. Qualquer versão razoável do modelo (2.10) envolverão no mínimo três parâmetros de covariância. (Diggle, Ribeiro e Christensen, 2003)

Ajustar uma função paramétrica ao variograma empírico é uma maneira de estimar os parâmetros para a covariância. Alguns métodos de ajuste como o de mínimos quadrados ponderados, o da máxima verossimilhança são usados para modelar o variograma.

O variograma empírico é necessariamente mais sensível a não especificação do valor médio da superfície $\mu(s)$. Especificamente, falha para ajustar a variação de amplitudes largas na resposta média induzirá a evidências falsas da correlação de grandes amplitudes em $Z(\cdot)$.

Estimação do variograma

O variograma (ou variograma empírico) é definido em (2.7) como

$$2\gamma(s_1 - s_2) \equiv \text{var}\left(Z(s_1) - Z(s_2)\right) \tag{2.23}$$

Assumiremos que processo é intrinsecamente estacionário, ou seja, satisfaz as seguintes condições $E\big(Z(s)\big) = \mu$, para todo $s \in \mathcal{D}$ e var $[Z(s_1) - Z(s_2)] = 2\gamma(s_1 - s_2)$, para todo $s_1, s_2 \in \mathcal{D}$.

Sobre a suposição de média constante (sem tendência) o estimador baseado no método dos momentos, referido também como estimador clássico, proposto por Matheron (1963) é

$$2\hat{\gamma}(h) \equiv \frac{1}{|N(h)|}\sum_{N(h)}(Z(s_i) - Z\big(s_j\big))^2 \tag{2.24}$$

onde

$$N(h) = \{\big(s_i, s_j\big): s_i - s_j = h; i, j = 1, \dots, n\} \tag{2.25}$$

e $|N(h)|$ é o número de pares diferentes em $N(h)$. Note que $N(-h) \neq N(h)$, embora a propriedade $2\hat{\gamma}(-h) = 2\hat{\gamma}(h)$ do variograma empírico seja preservada; sobre a suposição de estacionariedade o estimador clássico é não viesado para o variograma empírico. O variograma estimado em (2.24) mede a estrutura da dependência espacial e a variabilidade espacial dos dados sobre a região de estudo. Representa o valor médio da diferença entre dois valores amostrais observados ao quadrado.

Quando os dados estão distribuídos irregularmente no $\Re^d$, o estimador (2.24) é geralmente suavizado por

$$2\gamma^+(h(l)) \equiv ave\{\Big(Z(s_i) - Z\big(s_j\big)\Big)^2 : (i, j) \in N(h); h \in T\big(h(l)\big)\} \tag{2.26}$$

onde a região $T\big(h(l)\big)$ é a tolerância especificada na região $\Re^d$ sobre $h(l),\ l = 1, \dots, K$, e $ave\{\cdot\}$ denota uma possível média ponderada sobre os elementos em $\{\cdot\}$. As regiões de tolerância devem ser tão pequenas quanto possível para preservar a resolução espacial, mas grande o suficiente tal que o estimador $2\gamma^+$ seja estável (Cressie, 1993).

Muitas vezes as regiões $\{T\big(h(l)\big): l = 1, \dots, K\}$ são escolhidas para serem disjuntas; isto é análogo à suavização de um histograma de um conjunto de dados univariados. É natural então

pensar na estimação do variograma como sendo uma janela móvel, equivale ao estimador de densidade Kernel (Rosenbaltt, 1985).

É preferível estimar o variograma do que o covariograma, já que este último pode não ser estimado. O covariograma $C(h)$ ou o correlograma $\rho(h)$ podem ser estimados, quando $Z(\cdot)$ apresenta estacionariedade de segunda-ordem ambos são definidos, porém quando $Z(\cdot)$ apresenta estacionariedade intrínseca eles podem não existir. As propriedades e vantagens do variograma em relação ao covariograma e correlograma, fazem com que ele seja a ferramenta mais usada para expressar a dependência espacial dos dados.

O variograma é uma curva que representa o grau da continuidade espacial, é uma função crescente da distância h, seu crescimento mais ou menos rápido representa a influência de uma observação sobre a outra mais afastada.

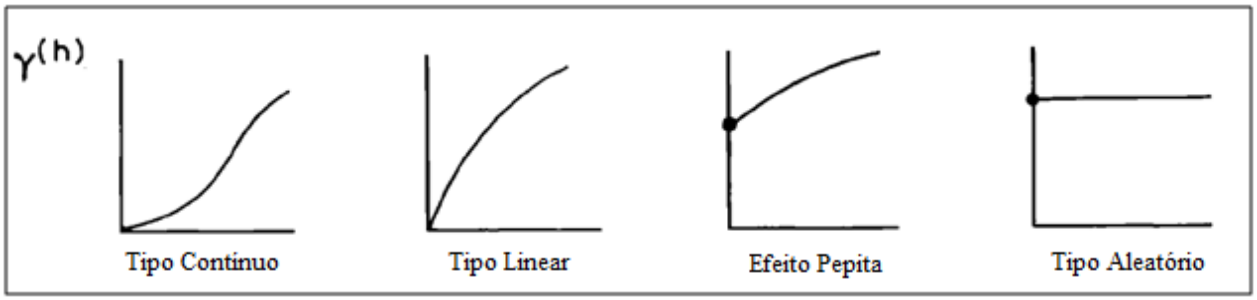

Figura 3 - Quatro formas diferentes para o variograma (Matheron, 1963).

Matheron (1963) descreve quatro tipos de comportamentos dos variograma (Figura3). O primeiro tipo tem uma forma parabólica próximo a origem e representa uma variável regionalizada com alta continuidade. O segundo tipo representa uma variável que tem continuidade espacial média. O terceiro tipo revela uma descontinuidade na origem e corresponde a uma variável apresentando não somente uma continuidade média, mas também um efeito pepita. O quarto tipo é um caso puramente aleatório, corresponde a uma variável aleatória. Entre esses quatro tipos aparecem inúmeros comportamentos intermediários para o variograma, nos quais está o objeto de estudo da geoestatística.

Modelos de variogramas teóricos

A condição necessária e suficiente para que uma família paramétrica de funções pertença à classe das funções de covariâncias é que seja definida positiva. O variograma empírico gerado a partir dos pares amostrais expressa a estrutura de covariância entre os dados deseja-se encontrar um modelo teórico que se ajuste a forma deste variograma o mais fiel possível.

Vários modelos paramétricos de variogramas são apresentados no Journel e Huijbregts (1978). Apresentaremos aqui apenas o modelo Gaussiano por ter sido o modelo adotado na modelagem geoestatística dos dados deste trabalho. Alguns modelos podem ser encontrados em Cressie (1993, p.61-63).

Modelo Gaussiano, válido no $\Re^d$, $d \geq 1$

$$\gamma(h;\, \theta) = \begin{cases} 0, & h = 0, \\ c_0 + c_g\left(1 - e^{-\left(\frac{\|h\|}{a_g}\right)^2}\right), & h \neq 0. \end{cases} \qquad (2.27)$$

$\theta = (c_0, c_g, a_g)$, todos maiores ou igual a zero, onde c_0 é o efeito pepita (*effect nugget*), c_g é o patamar (*sill*) e a_g é o alcance (*range*).

A Figura 4 ilustra a forma do ajuste do modelo Gaussiano aos pontos de um semivariograma. É um modelo que se ajusta bem em situações que apresentam grande continuidade. O efeito pepita para a figura apresenta descontinuidade para $h = 0$, patamar é a parte do semivariograma onde os dados não apresentam mais dependência espacial, e o alcance corresponde à distância da origem até o patamar.

O modelo Gaussiano é classificado como modelos com patamar. Pode ser instável sem o efeito pepita e tem a forma da distribuição normal acumulada.

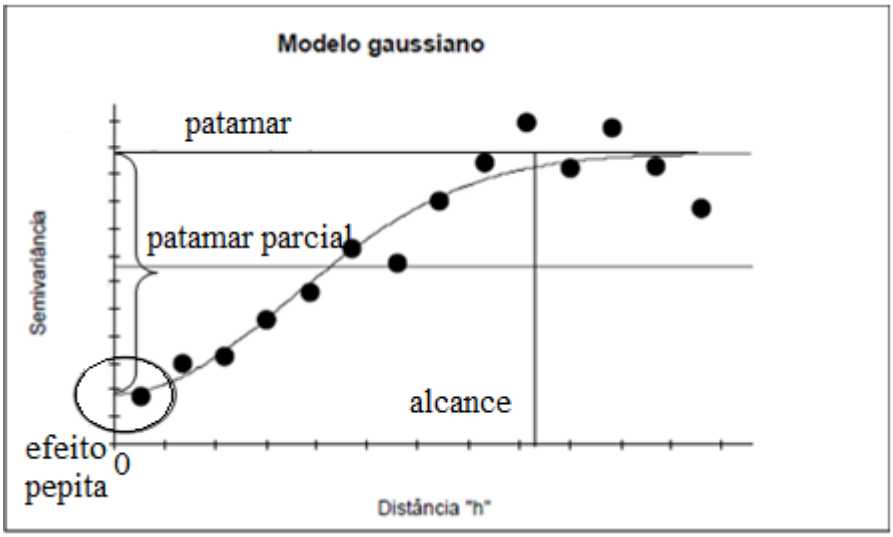

Figura 4 - Modelo Gaussiano ajustado a um semivariograma.

Métodos de estimação paramétrica para a estrutura de covariância espacial

A seguir são apresentados alguns métodos paramétricos propostos para ajustar um modelo teórico ao variograma empírico.

Método dos Mínimos Quadrados Ordinários

Este método especifica que θ é estimado minimizando o termo

$$\sum_{j=1}^{K}\{2\gamma^{\#}(h(j)e) - 2\gamma(h(j)e;\ \theta)\}^2 \qquad (2.28)$$

para alguma direção e. Múltiplas direções também poderiam ser calculadas para (2.28) adicionando-se as diferenças quadráticas apropriadas.

Método dos Mínimos Quadrados Generalizados

Suponha que o estimador do variograma $2\gamma^{\#}$ seja obtido nos K *lags* $h(1), \ldots, h(K)$, onde K é fixado e a quantidade de dados que contribuem para o estimador em cada *lag* seja grande (no mínimo 30 pares de acordo com o Journel e Huijbregts (1978, p.194). Seja $\gamma(h;\ \theta)$ um modelo teórico cuja forma exata é conhecida exceto para o parâmetro desconhecido θ.

O método dos Mínimos Quadrados Ordinários é um procedimento puramente numérico que tem um a interpretação geométrica atrativa. Para preservar a geometria, e também introduzir

os conceitos de variação no procedimento, considere o critério de Mínimos Quadrados Generalizados. Suponha $2\gamma^{\#} \equiv (2\gamma^{\#}(h(1)), \ldots, 2\gamma^{\#}(h(K)))'$, um vetor $K \times 1$ de variáveis aleatórias, tem uma matriz de variâncias var $(2\gamma^{\#}) = V$, que pode depender de θ. Então o valor de θ que minimiza

$$\left(2\gamma^{\#} - 2\gamma(\theta)\right)' V^{-1}\left(2\gamma^{\#} - 2\gamma(\theta)\right) \tag{2.29}$$

onde $2\gamma(\theta) \equiv \left(2\gamma(h(1)); \theta), \ldots, 2\gamma(h(K))\right)'$, é o modelo teórico avaliados nos *lags* $h(1), \ldots, h(K)$ chamado estimador $\theta_V^{\#}$.

Os métodos dos Mínimos Quadrados Generalizados usam a estrutura de segunda-ordem para estimar o variograma e não faz suposições sobre a distribuição dos dados (Cressie, 1993, p.95).

Método dos Mínimos Quadrados Ponderados

Juntamente com o estimador de Mínimos Quadrados Ordinários $\theta_I^{\#}$ e o estimador de Mínimos Quadrados Generalizados $\theta_V^{\#}$ existe o estimador de Mínimos Quadrados Ponderados $\theta_\Delta^{\#}$, onde

$$\Delta \equiv \text{diag } \{\text{var } (2\gamma^{\#}(h(1))), \ldots, \text{var } (2\gamma^{\#}(h(K)))\} \tag{2.30}$$

É uma matriz diagonal $K \times K$ com variâncias especificadas ao longo da diagonal.

Método da Máxima Verossimilhança

Os procedimentos de estimação por Máxima Verossimilhança (ML) e Máxima Verossimilhança Restrita (REML) dependem da suposição de que o processo é Gaussiano. O problema com a estimação por ML é que os estimadores de θ são viesados, sendo proibitivo seu uso para amostras pequenas ou moderadas (Matheron, 1971; Mardia e Marshal, 1984). Um caso simples é quando os dados Z são de fato Gaussiano multivariado independentes, Gau $(X\beta, \sigma^2 I)$, produzindo um parâmetro $\theta = \sigma^2$ com pequena escala de variação. O estimador de Máxima Verossimilhança é $\hat{\sigma}^2 = \sum_{i=1}^{n} (Z(s_i) - X\hat{\beta})^2/n$, onde $\hat{\beta}$ é o estimador de Mínimos Quadrados Ordinários do vetor β de dimensão $q \times 1$. Sabe-se que o estimador $\hat{\sigma}^2$ é viesado, para torná-lo

adequado acrescenta-se o fator de correção $(n/(n-q))$, assim $(n/(n-q))\hat{\sigma}^2$ é um estimador não viesado.

De forma mais geral, suponha que os dados são Gaussianos multivariado, Gau $(X\beta, \sum(\theta))$ onde X é uma matriz $n \times q$, com $q < n$, e que a matriz $n \times n$ $\sum(\theta) = (\text{cov}(Z(s_i), Z(s_j)))$ depende de θ através de $P = \{2\gamma: 2\gamma(\cdot) = 2\gamma(\cdot\,;\,\theta);\ \theta \in \ominus\}$. Então o "log" negativo da máxima verossimilhança é

$$L(\beta,\theta) = \left(\frac{n}{2}\right)\log(2\pi) + \left(\frac{1}{2}\right) log|\sum \theta| + \left(\frac{1}{2}\right)(Z - X\beta)' \sum \theta^{-1} (z - X\beta)\ ,\ \beta \in \Re^q, \theta \in \ominus \quad (2.31)$$

e os estimadores de Máxima Verossimilhança$\hat{\beta}$ e $\hat{\theta}$ satisfazem

$$L(\hat{\beta},\hat{\theta}) = \text{inf}\ \{L(\beta,\theta): \beta \in \Re^q, \theta \in \ominus\} \quad (2.32)$$

Método da Máxima Verossimilhança Restrita

O método da Máxima Verossimilhança Restrita, desenvolvido por Patterson e Thompson (1974), corresponde a minimizar a expressão abaixo para estimar θ.

$$L_W(\beta,\theta) = \left(\frac{n-1}{2}\right)\log(2\pi) + \left(\frac{1}{2}\right) log|A' \sum \theta\, A| + \left(\frac{1}{2}\right)(W - A'X\beta)'(A' \sum \theta\, A)^{-1}(W - A'X\beta) \quad (2.33)$$

Onde $A = (a_{ij})$ é uma matriz $(n-1) \times n$ cujos elementos são

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

Assumindo que o processo $Z(\cdot)$ tem média constante μ, então $A'X\beta = 0$ e consequentemente L_W não depende de β. De acordo com Diggle e Ribeiro (2000), o estimador de Máxima Verossimilhança Restrita é menos viesado para amostras pequenas.

Os métodos que utilizam a função de verossimilhança consideram toda a informação amostral decorrente do cálculo de todos os pares para todas as direções. Utilizam o variograma *cloud*, gráfico que mostra as contribuições individuais dos pares de pontos, é uma nuvem de pontos calculados para todas as distâncias e todas as direções.

Validação Cruzada (*Cross-Validation*)

Suponha que o modelo teórico para o variograma $2\gamma(h;\,\hat{\theta})$, $h \in \Re^d$, foi ajustado aos dados $\{Z(s_i): i = 1, \ldots, n\}$. Uma forma de diagnosticar algum problema com o ajuste obtido é fazer a validação cruzada para o modelo teórico.

A ideia é remover os dados um a um e usar o restante dos dados para predizer essa observação que foi removida. Então o erro de predição pode ser inferido dos valores preditos menos os valores observados. Isso auxilia na avaliação da variabilidade do erro de predição. O preditor $\hat{Z}(s_0)$, baseado no variograma ajustado e nos dados $\{Z(s_i): i = 1, \ldots, n\}$, do valor de $Z(\cdot)$ na localização $s_0 \in \mathcal{D}$ é conhecido, juntamente com a medida de erro quadrático médio $\sigma_k^2(s_0)$.

Se o modelo teórico para o variograma descreve adequadamente a dependência espacial implícita nos dados, então o valor de $\hat{Z}(s_0)$ deve ser próximo ao verdadeiro valor na amostra $Z(s_0)$. Analisar os resíduos padronizados preditos é uma forma de diagnosticar o ajuste do modelo para o variograma.

Na validação cruzada não é só o modelo que está sendo avaliado, mas também toda a modelagem do processo estocástico (a estacionariedade, isotropia, estimadores).

Método de Predição Espacial

Predição espacial refere-se à predição de $g(Z(\cdot))$ ou $g(W(\cdot))$ para os dados $Z(s_1), \ldots, Z(s_n)$ observados nas localizações conhecidas $s_1, \ldots, s_n$, o objetivo é inferir sobre quantidades aleatórias.

Krigagem é um método do erro quadrático médio mínimo de predição espacial que, geralmente, depende das propriedades de segunda-ordem do processo $Z(\cdot)$. Matheron (1963) chamou este método de predição espacial linear ótima, após os trabalhos de D. G. Krige na África do Sul, entretanto a formulação para a predição linear ótima não vieram dos trabalhos de Krige. As contribuições de Wold (1938), Kolmogorov (1941) e Wiener (1949) continham as equações de predição linear ótima que refletiam a ideia de que observações próximas ao ponto de predição deveriam obter uma ponderação maior na predição. Porém autores da área de meteorologia inseriram a predição linear ótima no contexto espacial em termos do variograma, usando a terminologia interpolação ótima em vez de krigagem.

Considerando que o modelo teórico para o variograma empírico é adequado e descreve a estrutura de dependência espacial dos dados de modo fidedigno, o próximo passo é fazer a predição de valores em locais não observados, para entender como o fenômeno ocorre fora da região amostrada. Alguns tipos de krigagem podem ser encontrados em Diggle e Ribeiro (2007) e Cressie (1993), aqui será abordada somente a Krigagem Ordinária, já que foi o método usado para os dados de salinidade.

A palavra krigagem é sinônimo de predição ótima, refere-se a fazer inferência de valores não observados do processo aleatório $Z(\cdot)$ dada por $\{Z(s): s \in \mathcal{D} \in \Re^d$ ou de $W(\cdot)$ para os dados

$$Z \equiv (Z(s_1), \dots, Z(s_n))' \tag{2.34}$$

observados nas localizações espaciais $s_1, \dots, s_n$.

Assumindo que nossos dados $Z = Z(s_1), \dots, Z(s_n)$ são gerados por um modelo gaussiano estacionário, escreveremos $W = W(s_1), \dots, W(s_n)$ para os valores não observados do sinal nas localizações amostradas $s_1, \dots, s_n$. Então W é gaussiano multivariado com vetor de média $\mu 1$, onde 1 denota um vetor cujos elementos são todos 1, e matriz de variância $\sigma^2 R$, onde R é a matriz $n \times n$ com elementos $r_{ij} = \rho(\|s_i - s_j\|)$. Similarmente, Z é gaussiano multivariado com vetor de média e matriz de variância (Diggle e Ribeiro, 2007)

$$\sigma^2 V = \sigma^2(R + \nu^2 I) = \sigma^2 R + \tau^2 I \tag{2.35}$$

Na Krigagem Ordinária o valor médio é tratado como desconhecido, enquanto que os parâmetros de covariância são assumidos conhecidos. Então o preditor é do tipo

$$\hat{T} = \mu + r' V^{-1}(Y - \mu 1) \tag{2.36}$$

o parâmetro μ é substituído pelo seu estimador de Mínimos Quadrados Generalizado,

$$\hat{\mu} = (1' V^{-1} 1)^{-1} 1' V^{-1} Y \tag{2.37}$$

com V dado por (2.35).

A predição pelo método da Krigagem Ordinária pode ser expressa como uma combinação linear $\widehat{W}(s) = \sum a_i(s)\, Z_i$. O termo a_i é chamado de predição ponderada, ou krigagem ponderada, e tem a propriedade $\sum a_i(s) = 1$ para qualquer localização. Queremos predizer o valor do sinal, $T = W(s)$, em qualquer localização s, usando os dados observados $Z = Z(s_1), \dots, Z(s_n)$, onde cada Z_i representa uma possível versão ruído correspondente a $W(s_i)$.

3. RESULTADOS

O estudo da salinização do solo foi desenvolvido pelos pesquisadores (Felipe Carmona, Ibanor Anghinoni e Eliseu José Weber) com o objetivo de mapear a ocorrência desse fenômeno nas planícies costeiras do Rio Grande do Sul. Segundo eles o levantamento das áreas afetadas ajuda a dimensionar o problema e a estabelecer estratégias de ação para diminuir ou corrigir sua ocorrência. O delineamento amostral foi definido pelos autores da pesquisa enquanto que a coleta dos dados ocorreu entre os meses de julho e setembro de 2008 e maio e agosto de 2009. A amostragem se estendeu pelos municípios de Rio Grande, localizado às margens do extremo sul da Laguna dos Patos, e Torres, último município do Litoral Norte do Estado, na divisa com Santa Catarina (Figura 5).

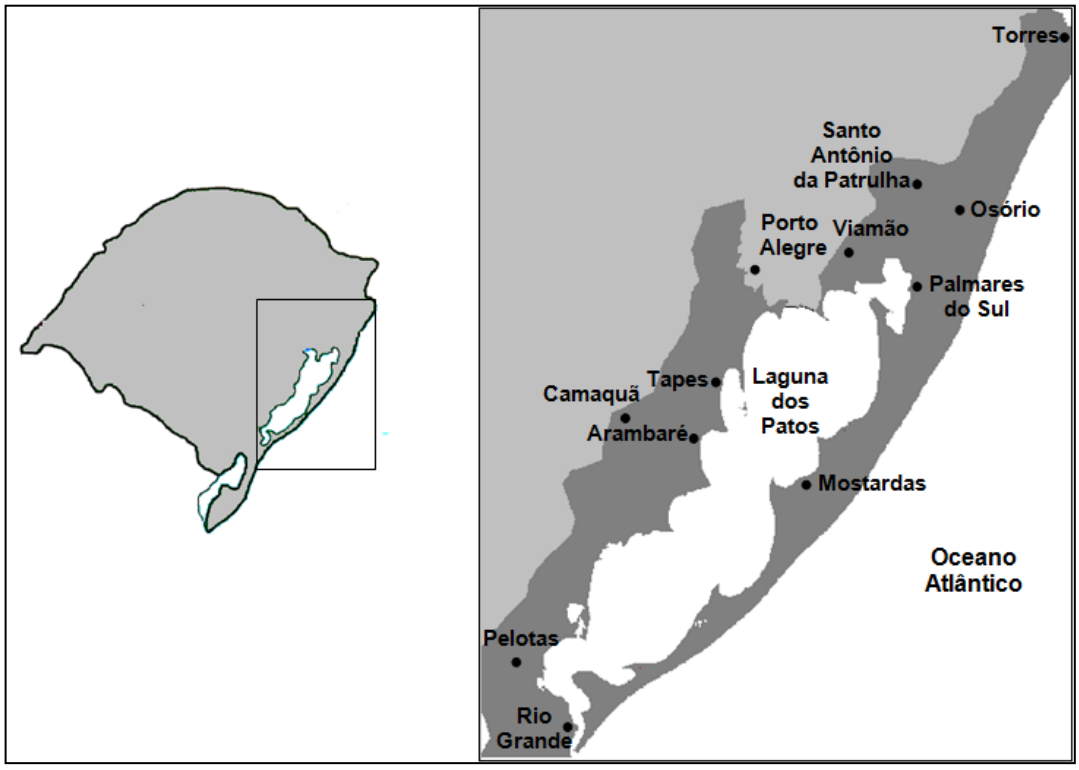

Figura 5 - Localização da área de estudo no Rio Grande do Sul, sendo a região escurecida os locais onde ocorreu a coleta de dados de solo.

As áreas de coleta ocorreram nos locais com histórico de cultivo de arroz irrigado por inundação ou que tinham potencial de uso para o cultivo. A amostra de solo é composta por 766 observações georreferenciadas com aparelhos GPS, destas foram determinadas os valores para as variáveis químicas que caracterizam a salinidade: condutividade elétrica do extrato de saturação (CE), sódio trocável (Na) e percentagem de sódio trocável (PST).

Inicialmente foram realizadas a análise exploratória dos dados através das medidas descritivas para cada uma das variáveis na região de estudo, tais como medidas de posição e dispersão, e gráficos como histograma e *box-plot* que fornecem informações sobre a distribuição dos dados e valores extremos. Posteriormente, a análise espacial foi desenvolvida visando identificar a dependência espacial dos dados na região de estudo, representada pelos valores das estimativas dos parâmetros de um modelo estatístico, que será usado para fazer predições em locais não amostrados através de um método de interpolação para gerar uma superfície contínua que descreva o comportamento da salinidade na área estudada. Dentre as três sub-divisões da Estatística Espacial, definido por Cressie (1993), a geoestatística será a metodologia adotada para a análise destas três variáveis.

O pacote *geo*R do software livre R foi usado para as análises estatísticas dos dados geoestatísticos. Os comandos, tutoriais e informações a respeito deste pacote se encontram no site http://www.leg.ufpr.br/geoR/, da Universidade Federal do Paraná. Para a geração de mapas foi usado o software *Idrisi Taiga* disponibilizado pela UFRGS através do Instituto de Ecologia.

Os resultados apresentados a seguir referem-se a cada uma das variáveis que caracterizam a salinidade (PST, Na e CE). No anexo desta monografia estão os comandos usados no pacote *geo*R para a análise dos dados georreferenciados.

Uma estratégia de análise adotada para este estudo foi dividir a região pesquisada em dois blocos, ou seja, lado oeste e lado leste da Laguna, já que o lado leste tem contato direto com o mar, enquanto que o contato do lado oeste ocorre no extremo sul do Estado, o fluxo de entrada de água salina, nesse caso, acontece em maior quantidade quando o nível da Laguna está mais baixo ou quando ventos em direção ao norte ocorrem com mais intensidade. Os blocos estão identificados no mapa (Figura 6). Como a Laguna está localizada geograficamente no meio da região de estudo os resultados gerados, sem considerar essa divisão, seriam inferidos através de

um mapa como sendo uma região contínua para cultivo de arroz, inclusive para a Laguna. Como fazer predição para um local onde não é possível o cultivo de arroz?

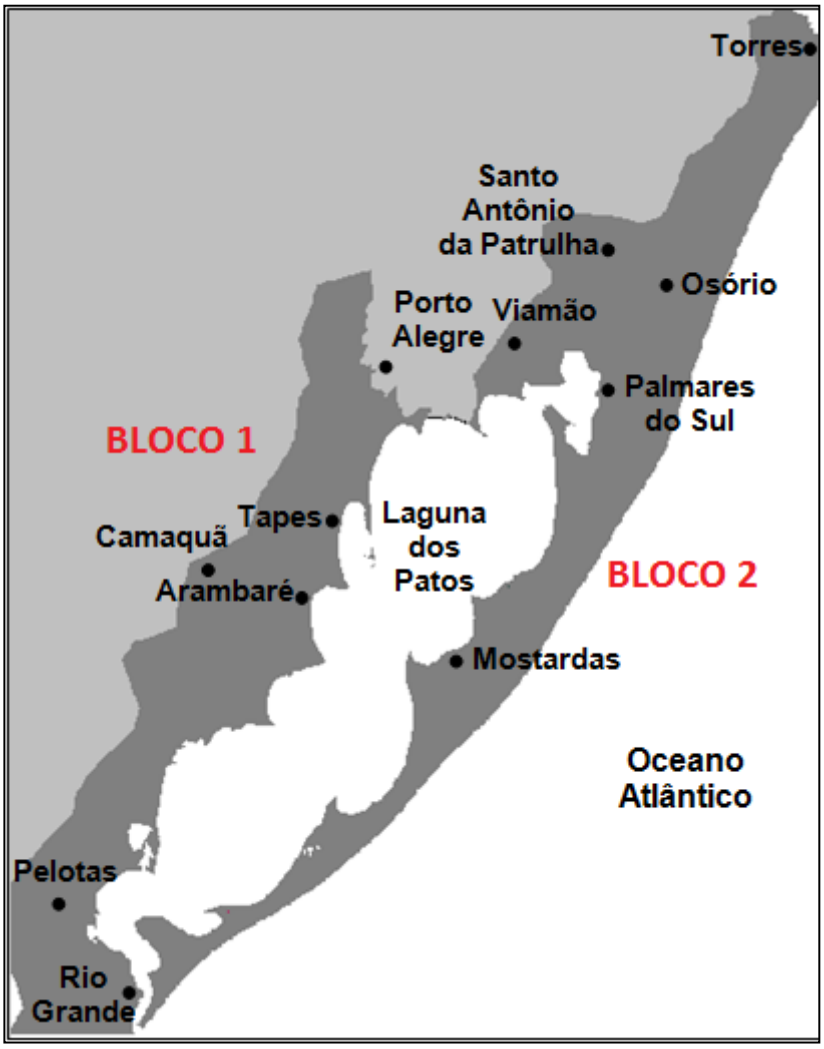

Figura 6 - Identificação do BLOCO 1 e BLOCO 2.

Os resultados apresentados nesta monografia referem-se apenas ao BLOCO 1, o objetivo é mostrar os passos da modelagem geoestatística realizada sob enfoque clássico para os dados espacialmente contínuos. Como este é um trabalho de modelagem feito para auxiliar os pesquisadores no entendimento da salinização na região baseado nos dados coletados, os resultados completos e finais não serão apresentados aqui. O procedimento da análise será o

mesmo adotado para o outro lado, BLOCO 2, porém com as adaptações necessárias para o contexto daquela região de estudo.

Em relação ao BLOCO 1, uma questão surge ao observar a Figura 6: a entrada do mar no extremo sul interfere na salinidade na região? Na medida em que a Laguna se distancia dessa entrada valores menores são observados para a variável estudada? A inclusão de uma covariável no modelo é necessária? A partir das perguntas acima, será investigada a possível existência da covariável distância da margem da Laguna. A significância estatística dessa covariável será testada para cada uma das variáveis, se não for verificada, os dados serão modelados sem a sua inclusão.

Algumas informações referentes ao BLOCO 1 são apresentadas abaixo, são estendidas para as três variáveis estudadas.

BLOCO 1

Tabela 1 - Informações das coordenadas.

	Coordenadas	
	coordX	coordY
min	357250	6440250
max	475250	6678250
max-min	118000	238000

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

Amostra	
Número de observações	315
Menor distância entre pares	500
Maior distância entre pares	256632

Tabela 3 - Informações da covariável: distância da margem da Laguna.

Covariável	
Distância mínima da margem da Laguna	0
Distância máxima da margem da Laguna	29428

A Tabela 1 apresenta informações das coordenadas de X e Y, como o menor e maior valor da coordenada. O BLOCO 1, lado oeste da Laguna, possui 315 observações das 766 totais. A menor distância verificada entre pares de observações é de 500 m e a maior distância é de 256.632 m. Cada uma das 315 observações está a uma determinada distância da margem da Laguna, a Tabela 3 mostra que a menor distância é zero e a maior distância é 29.428 m.

3. 1. RESULTADOS VARIÁVEL PST

O valor da variável percentagem de sódio trocável (PST) é calculada pela equação PST = (Na^+/ CTC_{pH}) x 100, onde o denominador corresponde a soma dos cátions Na^+, K^+, Ca^{2+}, Mg^{2+} e H^++Al^{3+}.

A estatística descritiva é uma análise exploratória que resume a informação contida nos dados, organiza e os descreve através de gráficos, medidas de posição e dispersão, apresentando uma visão geral da sua variação. A Tabela 4 apresenta os resultados numéricos para as medidas de posição para a variável PST. Nota-se que a média e a mediana apresentam valores distantes indicando assimetria dos dados e grande amplitude. Para verificar a distribuição dos dados foi construído o histograma (Figura 7) e pela análise visual a assimetria dos dados é verificada, assimétricos à esquerda. Sobre as colunas do histograma encontram-se a quantidade de valores observados naquele intervalo, tem-se um valor mais afastado dos outros, refere-se ao valor de máximo da Tabela 4, possivelmente seja um valor *outlier*.

Tabela 4 - Medidas Resumo dados originais.

PST			
Min.	Mediana	Média	Max.
0.10	1.60	2.597	30.10

Para contornar essa forte assimetria decidiu-se fazer uma transformação nos dados do tipo ln (PST+1). A Tabela 5 apresenta os resultados numéricos para os dados transformados, nitidamente ocorreu uma aproximação da mediana com a média, e a amplitude entre os valores máximos e mínimos diminuiu. Porém, a mudança na distribuição dos dados é confirmada pelo histograma da Figura 7 (à direita), a assimetria continua existindo mas a distribuição das observações sobre o gráfico não é tão discrepante como verificada antes.

Tabela 5 - Medidas Resumo dados transformados.

LogPST			
Min.	Mediana	Média	Max.
0.09531	0.9555	1.0720	3.4370

Um gráfico mais informativo quanto a posição, dispersão, assimetria e valores extremos é o *box-plot*, Figura 8. Representa os dados por um retângulo construído com base nos quartis (divide os dados em três partes iguais 0.25, 0.5, 0.75) e pelas linhas que vão do retângulo até os valores mais atípicos. Para os dados originais o gráfico mostra valores bem afastados da mediana (linha no centro do retângulo), já para os dados transformados essa dispersão diminui, apesar de ainda ser visualmente perceptível alguns valores atípicos para os valores observados. A partir deste ponto prosseguiremos as análises com os dados transformados (LogPST).

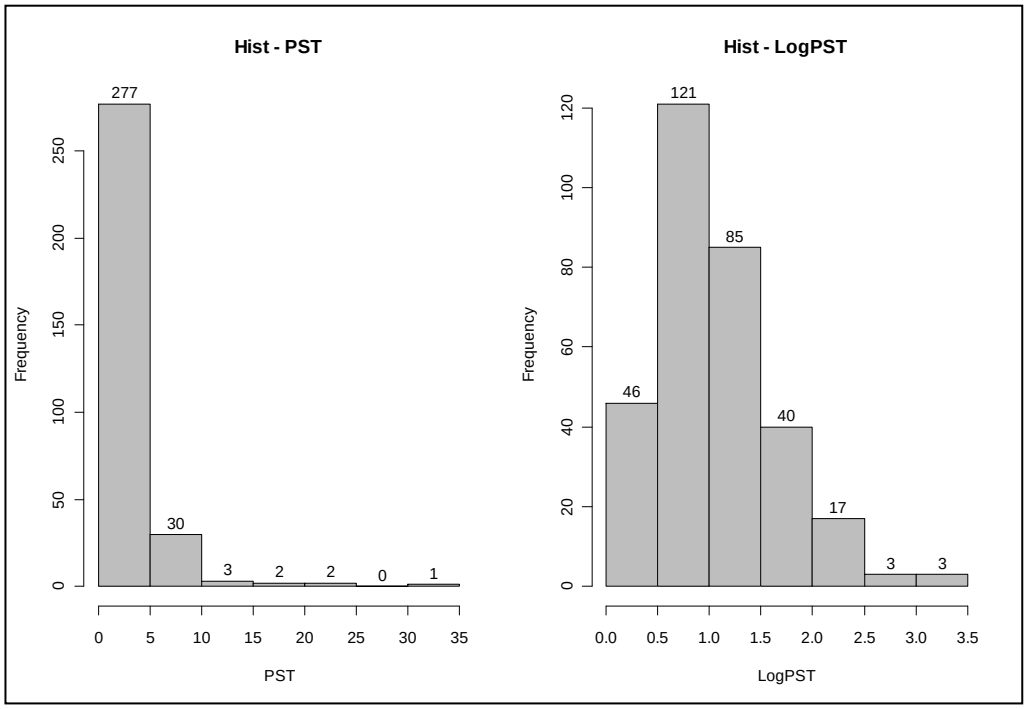

Figura 7 - Distribuição dos dados originais (à esquerda) e distribuição dos dados transformados (à direita), ambos representados pelo histograma.

O primeiro passo na análise exploratória espacial é explorar os valores observados considerando as suas localizações geográficas. No programa *geo*R a função *points.geodata* produz um gráfico com estas características (Figura 9) que ajuda na inspeção de possíveis valores discrepantes com relação a sua vizinhança espacial. Espera-se que observações geograficamente próximas tenham valores mais similares.

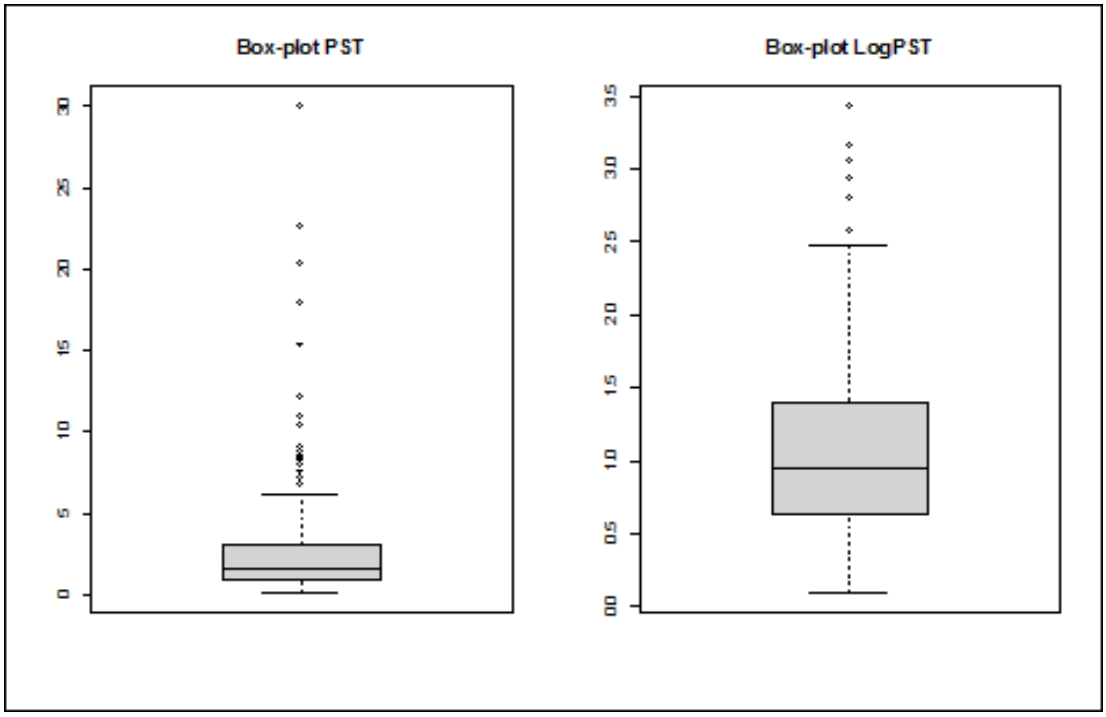

Figura 8 - Box-plot para os dados originas (à esquerda) e para os dados tranformados (à direita).

A Figura 9 apresenta círculos de tamanhos e cores diferentes que indicam os valores dos quartis para os valores medidos, ou seja, apresentam os dados agrupados em quatro classes. Quanto maior o círculo maior é o valor observado da variável LogPST. Analisando visualmente a figura é possível localizar observações com valores acima dos observados para aquela vizinhança (círculo na parte superior da Figura 9, além de outros pontos que também apresentam essa característica), isso pode ocorrer devido ao uso corrente de produtos para o cultivo de arroz nesta região que pode ter contaminado o solo, ou por existir na região algum canal que seja muito salino. Outra característica verificada nessa figura refere-se ao padrão observado para os valores que tangem a margem da Laguna. Quanto mais próximo da margem maior é o valor medido para a variável indicando que pode haver uma tendência relacionada à margem da Laguna. Observando este comportamento decidiu-se modelar os dados de duas maneiras: com e sem a inclusão da covariável distância da margem da Laguna.

O banco de dados é composto pelas coordenadas X e Y, pelo valor da variável LogPST e pelas distâncias calculadas para cada observação quanto ao afastamento da margem da Laguna.

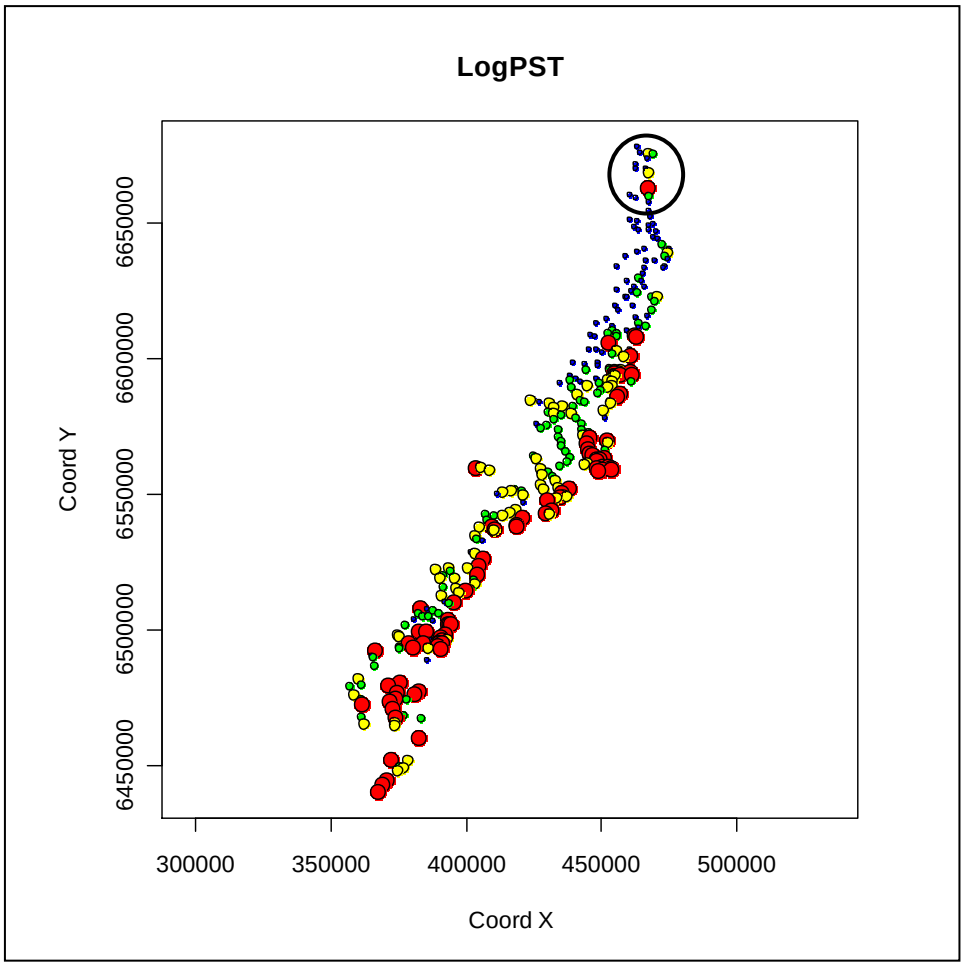

Figura 9 - Gráfico mostrando a localização dos 315 pontos amostrados para o LogPST.

A Figura 10 apresenta um gráfico 2x2 exploratório como outra forma de investigar o comportamento dos dados, no topo esquerdo os pontos localizados em suas respectivas coordenadas com cores e formas que correspondem aos quantis da distribuição dos dados são apresentados, no canto inferior direito tem a distribuição dos dados visualizados sobre o plano (coordenadas X e Y e o valor da variável) e a altura de cada ponto corresponde aos seus valores medidos. Os dois gráficos restantes referem-se à dispersão dos valores observados versus suas coordenadas separadamente.

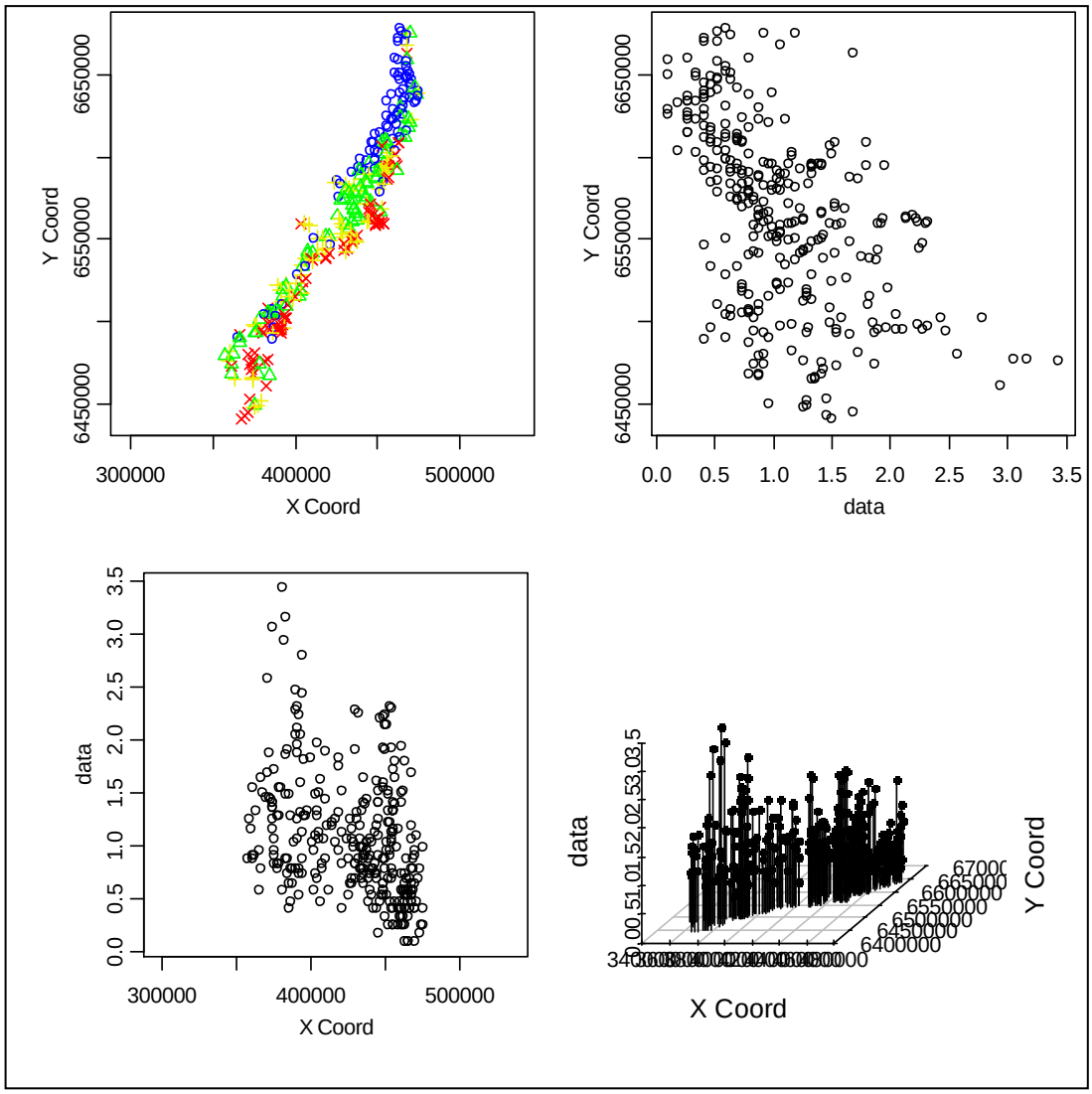

Figura 10 – Pontos localizados em suas coordenadas (topo esquerdo), valores de LogPST versus as coordenadas (topo direito e painel inferior esquerdo) e pontos dispersos sobre as coordenadas (painel inferior direito).

Para a avaliação da dependência espacial dos dados o variograma empírico é a ferramenta exploratória utilizada na geoestatística a fim de identificar a correlação das observações na medida em que aumenta a distância entre elas. Para um conjunto de dados geoestatísticos o variograma para os valores observados, se assumirmos que o processo gerador é um processo gaussiano estacionário, a correlação entre eles deve depender somente da distância, quanto maior a distância entre as observações a correlação tende a zero. O comportamento espacial para as observações da variável LogPST pode ser visto na Figura 11. No lado esquerdo é mostrado o

variograma para o modelo sem covariável e no direito o modelo com a covariável distância da margem da Laguna.

Aparentemente o variograma de nuvem (ou variograma *cloud*), gráfico de dispersão das distâncias entre duas observações para toda amostra versus o valor do semivariograma calculado para elas, para um modelo considerando a covariável parece ser menos disperso da nuvem de pontos em relação ao outro. Porém, a semivariância apresentada pelos *bins* (meio da Figura 11) para os dois casos não parecem ser muito diferentes na forma, apesar do caso com a covariável apresentar valor do patamar um pouco maior. O gráfico dos variogramas *bins* correspondem à média calculada para cada *lag* do variograma *cloud*. O *box-plot* para os *lags* (quantidade fixada na rotina de distâncias a serem mostradas no gráfico) assim como para o variograma *cloud*, parece ser mais estável para o caso da inclusão da covariável. O quanto essa diferença é significativa será avaliada mais pra frente.

O variograma *bin* é monótono e crescente, analisando as características do variograma deste tipo para essa variável, nota-se que na medida em que a distância aumenta entre as observações tem-se um crescimento quase linear da correlação espacial entre elas. Em ambos os casos, o efeito pepita está em torno de 0.1, a continuidade espacial (alcance) é crescente para os dois casos, porém os valores são diferentes e a variância do processo observado (patamar) parece ser menor para o caso de não inclusão da covariável (0.6), esta corresponde ao valor em que a dependência espacial é nula. O estimador para o variograma usado para os dois casos é o clássico, é baseado no método dos momentos.

O parâmetro *sill* é igual à variância dos dados e significa que não existe qualquer relação entre os pares de dados considerados estabilizando nesse valor. O crescimento do variograma até interceptar o patamar se dá pela função de correlação onde as características mais importantes a serem avaliadas são quando a distância está próxima de zero e o quão rapidamente a correlação aproxima-se de zero quando a distância cresce, refletindo desta forma, a correlação espacial do processo. Os resultados gráficos mostrados na Figura 11 foram gerados assumindo que o processo é isotrópico, ou seja, que o variograma gerado independe da direção dos dados.

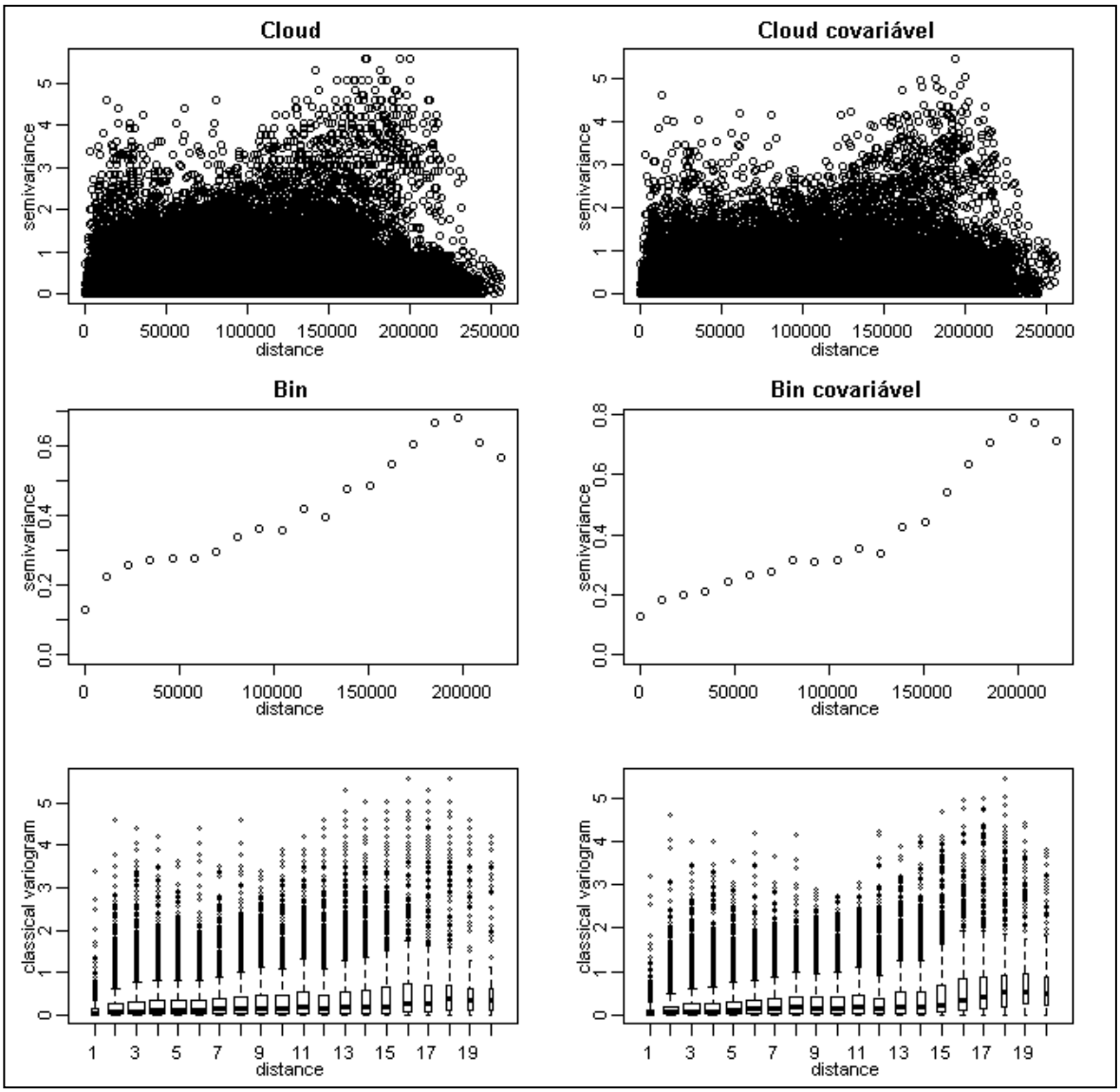

Figura 11 - Variograma cloud (superior), variograma bin (meio) e box-plot para os lags (inferior), sem a covariável (esquerda) e com a covariável (direita).

Tabela 6 - Resultados numéricos para os variogramas sem e com a covariável.

LogPST	**Bin**	**Omnidirectional**		
tendência constante (s/. cov)	var.mark	beta.ols		max.dist
	0.3392067	1.072384		225789.5
tendência distância (c/. cov)	var.mark	beta.ols		max.dist
	0.3392067	1.318012	-0.0000287	225789.5

A Tabela 6 apresenta os resultados para os variogramas relacionados à variância dos dados (var.mark), à máxima distância entre os pares de observações (max.dist) e aos parâmetros da parte média do modelo ajustado por mínimos quadrados ordinários (beta.ols).

Os variogramas não são usados apenas como ferramenta exploratória, mas também para estimar parâmetros, visto que é um estimador não viesado do variograma teórico. Ajustar uma função de covariância paramétrica ao variograma é uma maneira de estimar os parâmetros da estrutura de covariância espacial. O ajuste pode ser feito através de uma curva gerada por métodos de estimação como: método dos mínimos quadrados ordinários (OLS), mínimo quadrados ponderados (WLS), método da máxima verossimilhança (ML) e método da máxima verossimilhança restrita (REML); os dois últimos métodos se ajustam a todos os valores amostrados (variograma *cloud*), já os primeiros consideram somente os pontos do variograma *bin* para o ajuste. Esses métodos necessitam da especificação de um modelo paramétrico para a estrutura de covariância representada pelo variograma que permite gerar estimativas para os parâmetros do variograma teórico.

O modelo escolhido para representar a estrutura de covariância ajustado para os pontos do variograma dos dados da variável LogPST, com ou sem a covariável, foi o Gaussiano, porque visualmente a distribuição dos pontos no variograma *bin* (Figura 11) parece ser representada adequadamente pela forma do modelo Gaussiano.

A Figura 12 apresenta retas ajustadas adicionadas ao variograma pelos quatro métodos citados usando o modelo Gaussiano com e sem a inclusão da covariável. Nesse ajuste especifica-se o valor para o efeito pepita e os valores iniciais para os parâmetros de covariância espacial, patamar e alcance, usando a função *eyefit* do pacote *geo*R. Isso é um critério que pode ou não ser adotado na estratégia de modelagem, depende do pesquisador. Nessa etapa também

especificamos a covariável na rotina do programa. Observando os ajustes nos dois casos, percebe-se que os métodos OLS e WLS se ajustam melhor aos pontos do variograma empírico quando comparados com o ML e REML, porém o ajuste feito por esses últimos consideram a nuvem de pontos do variograma *cloud*, e não somente os pontos médios calculados para o variograma *bin*.

Verificaremos na validação cruzada o quanto essa disparidade visualizada para os ajustes das retas através dos métodos estão sendo diferentes.

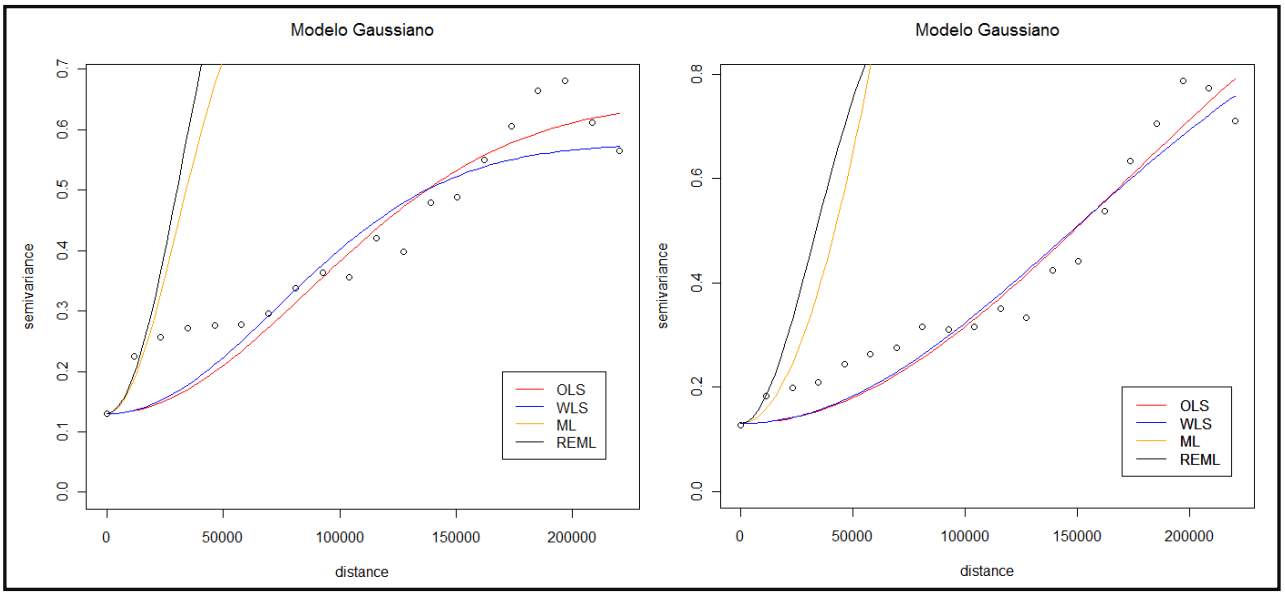

Figure 12 – Diferentes métodos de ajuste para o modelo Gaussiano sem a covariável(à esquerda) e com a covariável (à direita).

As Tabelas 8 e 9 apresentam os resultados numéricos para os ajustes mostrados acima, na modelagem fixou-se o valor do efeito pepita (*nugget effect*) e especificou-se os mesmos valores iniciais para os quatro métodos de ajustes fixando-se a inclusão ou não da covariável. Analisando a tabela para o caso da não inclusão da covariável, o método dos mínimos quadrados ponderados é o que apresenta menor valor para os parâmetros da estrutura de covariância. O método da máxima verossimilhança e o da máxima verossimilhança restrita apresentam valores para o AIC (*Akaike Information Criteria*) e o BIC (*Bayesian Information Criteria*) muito próximos, o que não auxilia na decisão de qual método utilizar para a predição

Antes de analisar os resultados da Tabela 9, verifica-se que a inclusão da covariável não é estatisticamente significativa visto que o intervalo com 95% de confiança contém o valor zero, indicando que não existem evidências significativas de que a covariável para esta variável seja relevante. Dessa forma, será feita a validação cruzada para os métodos de ajustes para o modelo Gaussiano que não considerou a inclusão da covariável.

As Figuras 13 e 14 mostram o resultado gráfico para a validação cruzada, na Tabela 7 são apresentadas as medidas resumo para os valores da diferença entre os dados e os valores preditos. A validação cruzada consiste em retirar a observação e usar o restante do conjunto de dados para predizer os vetores geográficos observados auxiliando na avaliação do ajuste obtido para o variograma empírico. Pelos resultados os métodos têm valores próximos para todas as medidas, apesar do método REML apresentar valores um pouco diferente em comparação aos outros métodos. Observando os resultados gráficos os métodos ML e REML parecem apresentar resultados mais adequados, pois os gráficos de dispersão da nuvem de pontos parecem estar mais próximos da reta, apesar dos *outliers*, e a distribuição dos dados menos os preditos avaliada pelo histograma e pelo gráfico da probabilidade observada versus a probabilidade teórica (círculo na Figura 14) está mais adequada para estes métodos.

Como numericamente não foi possível identificar um resultado diferente entre os métodos, apesar de graficamente os métodos ML e REML serem melhores, a krigagem (método de interpolação) será empregada para fazer a predição da superfície em locais não observados. Os quatro métodos ajustados para o modelo Gaussiano serão empregados para predizer valores para toda a malha descrevendo desta forma, a distribuição da salinidade caracterizada pela variável LogPST.

As Figuras 15 e 16 descrevem os resultados gráficos para a krigagem para os diferentes ajustes. Constata-se que os mapas gerados pelos métodos OLS e WLS são bem parecidos diferenciando-se apenas pela parte circulada na Figura 15 e levemente pela amplitude dos valores na escala de cores, sendo um pouco mais ampla pelo método WLS.

Observando os mapas gerados pelos métodos ML e REML, Figura 16, a distribuição contínua da superfície parece captar mais detalhes, com uma amplitude de valores (identificadas pela escala de cores) mais ampla em relação aos ajustes anteriores. Comparando um método com o outro, não parece existir diferença entre as superfícies.

Através dos mapas de superfícies e dos resultados verificados anteriormente, realizou-se a validação dos dados da variável LogPST sobre as superfícies geradas pelo método ML e REML. O objetivo foi auxiliar na escolha do método de estimação para o modelo Gaussiano para descrever a distribuição da salinidade na região. A justificativa para a escolha destes dois métodos decorre do fato de que eles consideram toda a informação dos pontos do variograma de nuvem resultando neste caso, em mapas que representam melhor o fenômeno. A Figura 17 mostra alguns pontos já verificados antes fora do padrão em relação a sua vizinhança (círculos na figura apontam essa disparidade) o que dificulta a validação. No entanto, visualmente não parece existir diferença entre os dois mapas.

Não parece existir diferença entre os dois métodos, porém o método REML apresentou maior amplitude de predição, representando com mais detalhes a superfície em relação ao outro método. Assim, o modelo Gaussiano ajustado pelo método de estimação REML foi escolhido para descrever a variabilidade espacial dos dados para a variável LogPST, e suas estimativas (Tabela 8 coluna 5) usadas para gerar uma superfície para toda a região de estudo representando a distribuição da salinidade.

Tabela 7 - Medidas para a diferença entre os dados e os preditos obtidos da validação cruzada.

Método de ajuste	Min.	Mediana	Média	Máx.	Desvio
OLS s/ cov	-1.268	-0.05118	0.0005924	1.768	0.4417976
WLS s/ cov	-1.262	-0.05057	0.0005824	1.778	0.4401874
ML s/ cov	-1.251	-0.04506	0.0005035	1.802	0.4335334
REML s/ cov	-1.616	-0.03005	0.0009945	1.596	0.4121591

Tabela 8 - Resultados numéricos para o ajuste sem a covariável.

Método de ajuste	OLS	WLS	ML	REML
nugget effect	0.13	0.13	0.13	0.13
partial sill	0.5168	0.4467	0.7896	1.114
range parameter	122386.3403	103481.8716	42950	47466
practical range	211828.5	179108.3	74339.09	82155.14
sum of squares	0.0641	209.4216	---	---
maximised log-likelihood	---	---	-186.9	-184.0
AIC	---	---	379.9	373.9
BIC	---	---	391.1	385.2

Tabela 9 - Resultados numéricos para o ajuste com a covariável.

Método de ajuste	OLS	WLS	ML	REML
nugget	0.13	0.13	0.13	0.13
partial sill	1.1568	0.9287	6.455	0.9059
range parameter	238801.05	206903.5367	172432	46350
practical range	413321.2	358112.4	298449.2	80230.38
sum of squares	0.0579	142.4191	---	---
maximised log-likelihood	---	---	-204.1	-182.5
covariável	---	---	-0.000013538	-0.0000098245
IC covariável	---	---	(-0.0002753; 0.00000045676)	(-0.000030919; 0.000011271)
AIC	---	---	416.2	373
BIC	---	---	431.3	388

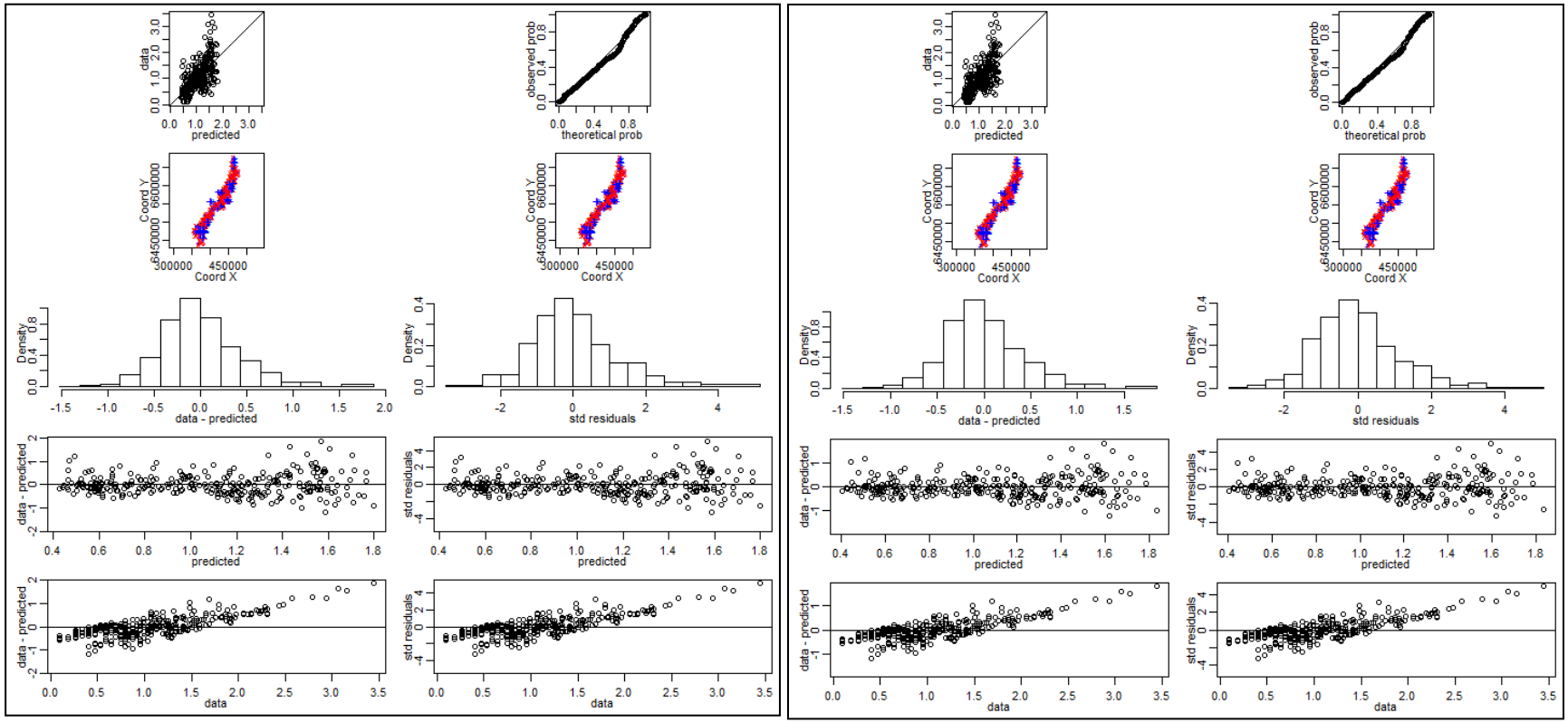

Figura 13 - Validação cruzada sem covariável: método OLS (à esquerda) e método WLS (à direita).

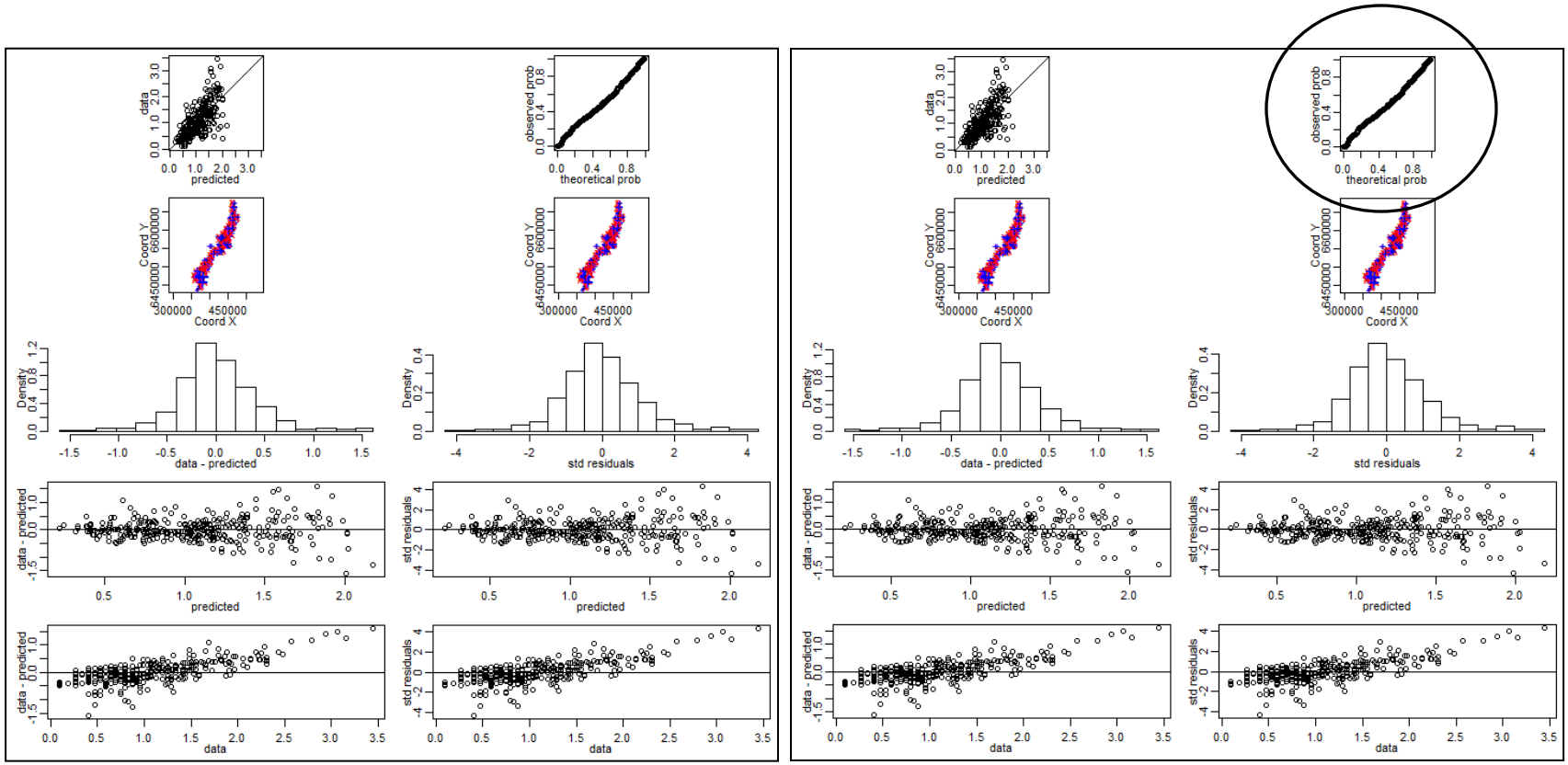

Figura 14 - Validação cruzada sem covariável: método ML (à esquerda) e método REML (à direita).

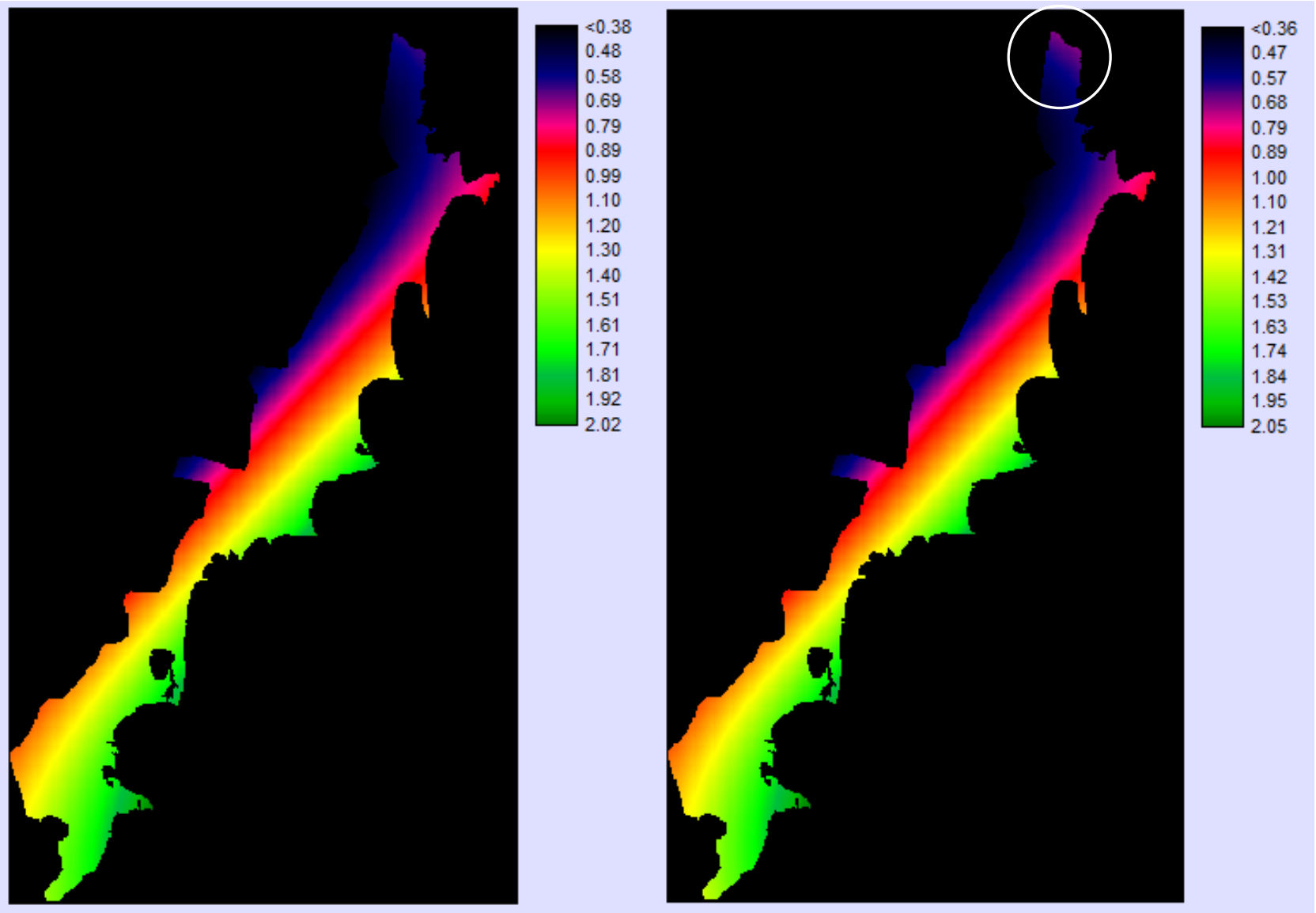

Figura 15 - Superfície gerada a partir dos parâmetros estimados do modelo Gaussiano ajustado pelo método OLS (à esquerda) e WLS (à direita).

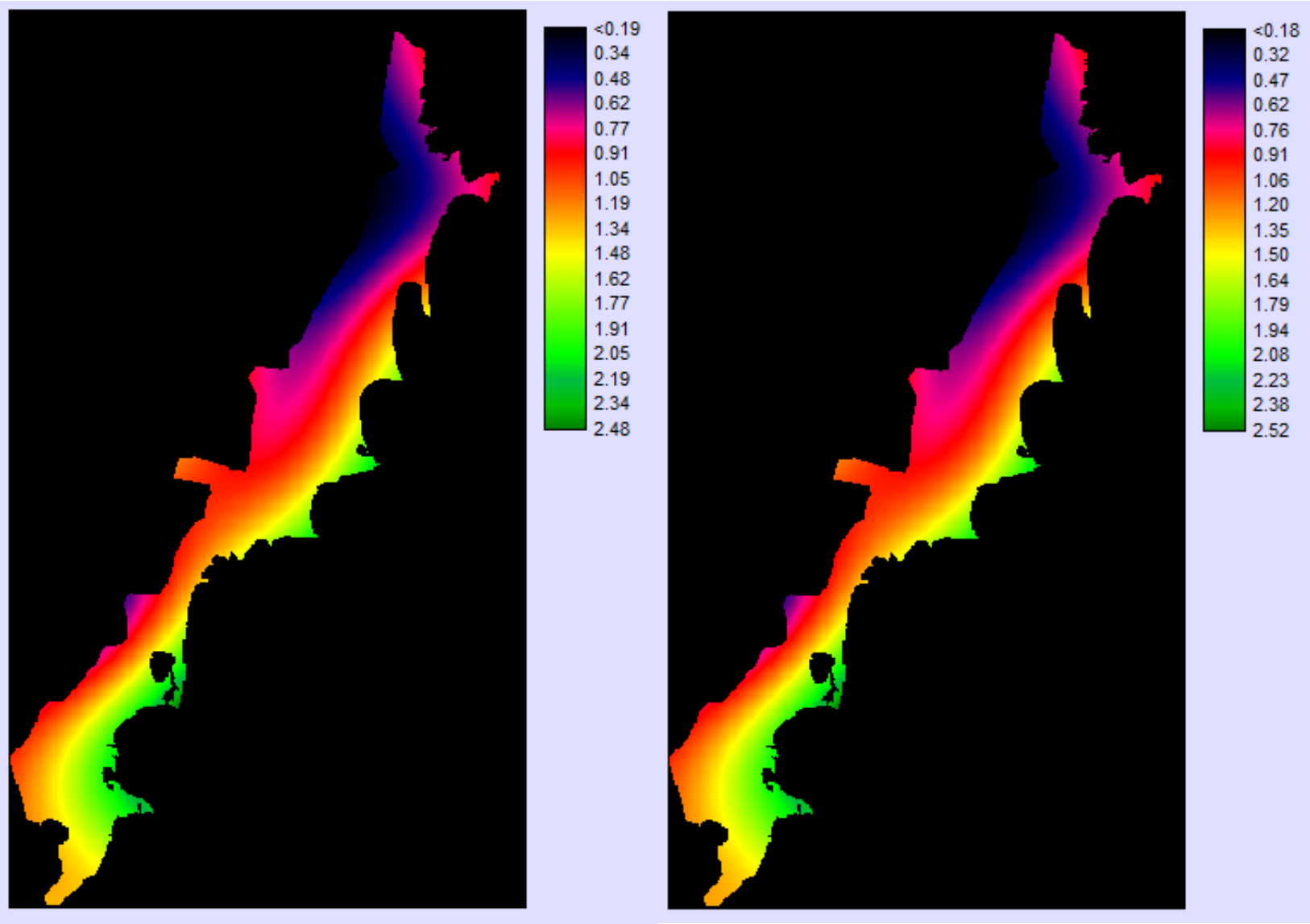

Figure 16 - Superfície gerada a partir dos parâmetros estimados do modelo Gaussiano ajustado pelo método ML (à esquerda) e REML (à direita).

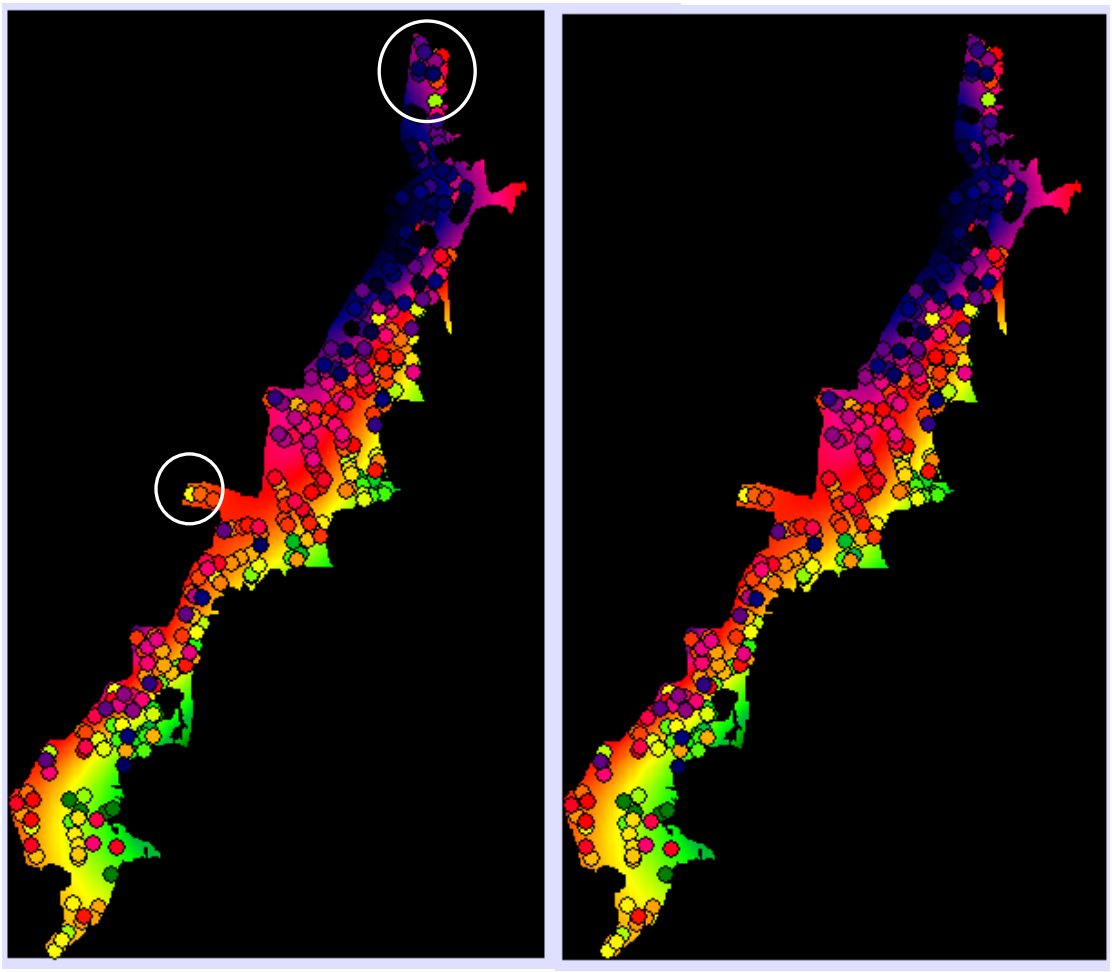

Figura 17 - Sobreposição dos dados no mapa de superfícies (ML à esquerda e REML à direita).

3. 2. RESULTADOS VARIÁVEL Na

O procedimento de análise para a variável sódio trocável (Na) é o mesmo adotado para a variável anterior. A análise exploratória para a variável Na é descrita pelas medidas resumo e é apresentada na Tabelas 10. Verifica-se uma grande amplitude entre o valor mínimo e máximo, a média e a mediana possuem valores distantes indicando assimetria nos dados. A Figura 18 mostra o histograma para os dados originais (à esquerda), a maior parte das observações possuem valores entre 0 e 200 para a variável. No entanto, identifica-se visualmente seis valores muito afastados caracterizando os *outliers*. Novamente uma transformação foi aplicada para o desenvolvimento da análise dos dados.

A transformação adotada para a variável Na é da forma ln (Na+1), chamaremos de LogNa. A alteração no comportamento dos dados muda significativamente, a Tabela 11 apresenta resultados bem diferentes do verificado anteriormente. A distribuição dos dados está mais simétrica. O *box-plot* (Figura 19) auxilia na identificação dos *outliers*, mesmo após a transformação seis valores continuam sendo atípicos em relação ao restante dos valores amostrais.

Tabela 10 - Medidas resumo para os dados originais.

Na			
Min.	Mediana	Média	Max.
2.00	37.00	67.38	1286.00

Tabela 11 - Medidas resumo para os dados originais.

LogNa			
Min.	Mediana	Média	Max.
1.099	3.638	3.711	7.160

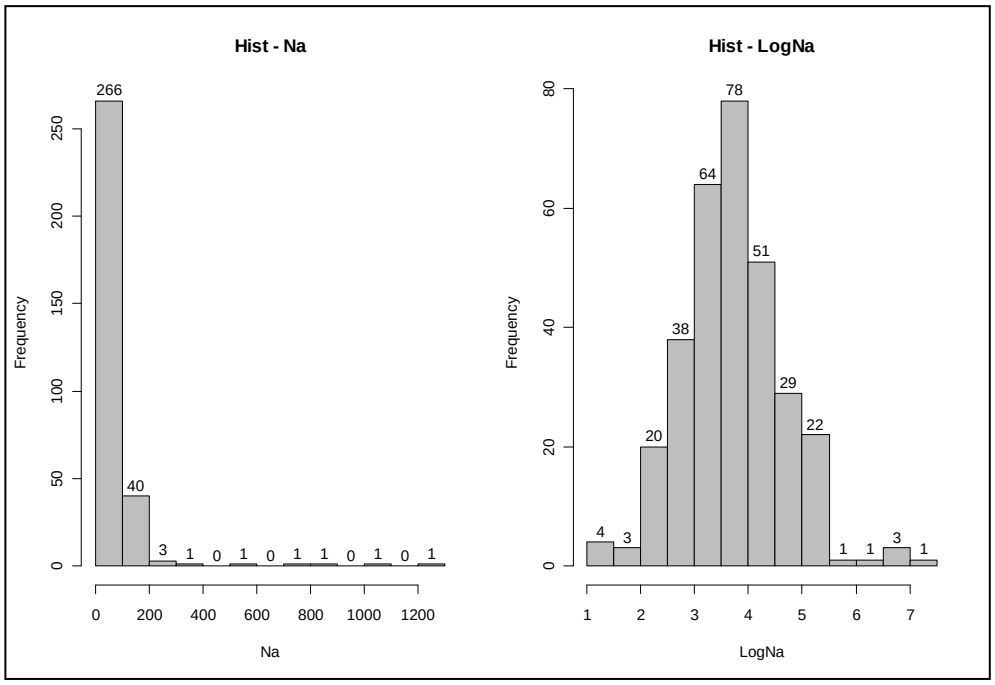

Figura 18 - Histograma para os dados originais (à esquerda) e os dados transformados (à direita).

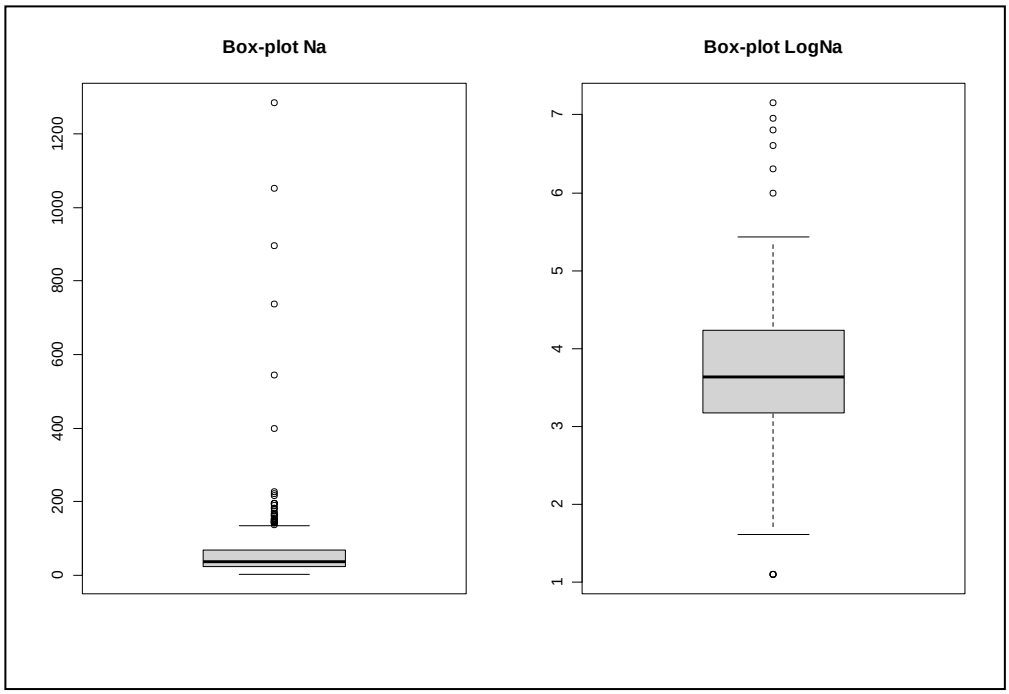

Figura 19 - Box-plot para os dados originais (à esquerda) e os dados transformados (à direita).

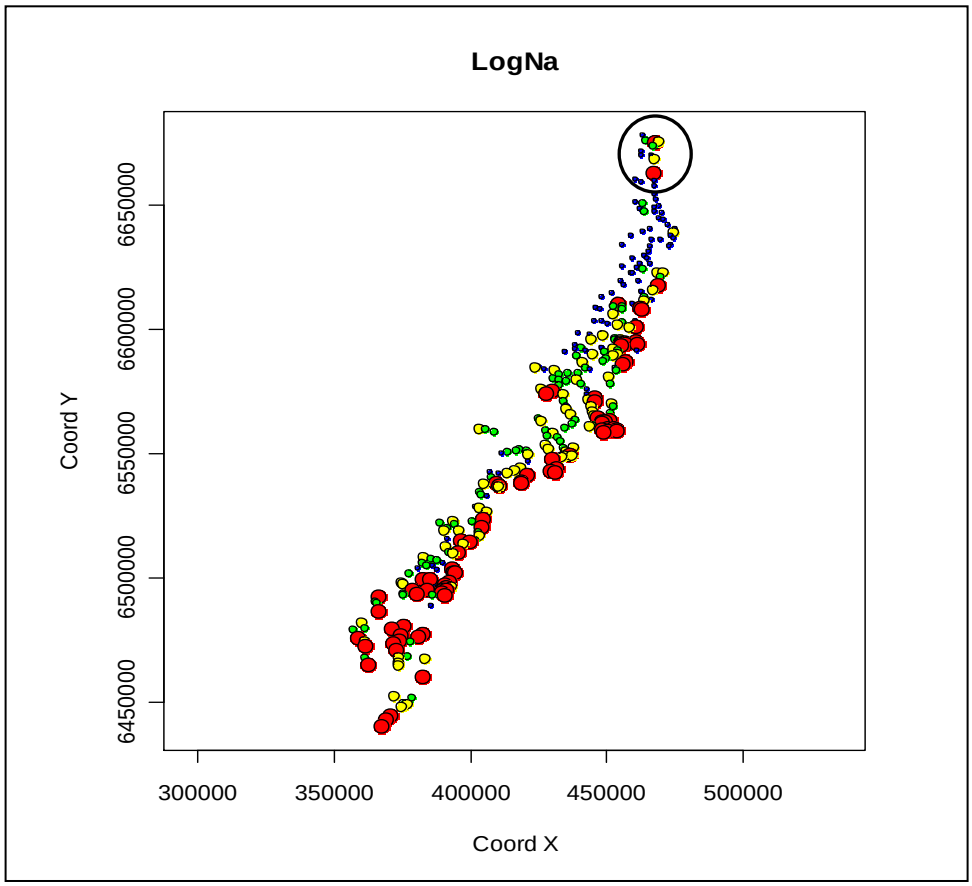

Figura 20 - Gráfico mostrando a localização dos 315 pontos amostrados para o LogNa.

A disposição dos dados para a variável LogNa é vista na Figura 20. Os maiores valores estão concentrados na margem da Laguna, enquanto que na parte superior da figura (circulada), e em outros locais do gráfico, valores diferentes em relação à vizinhança são identificados. Em decorrência do padrão percebido ao longo da margem a inclusão da covariável distância da margem da Laguna será avaliada no modelo.

Analisando visualmente a Figura 21, nota-se uma dispersão maior dos dados em relação à coordenada Y, e concentração no centro em relação a coordenada X. Os pontos dispersos sobre as coordenadas (painel inferior direito) mostram a distribuição dos dados no plano onde a altura de cada ponto corresponde ao seu valor. A distribuição dos valores parece seguir um padrão na medida em que se afasta da margem da Laguna (topo esquerdo da Figura 21).

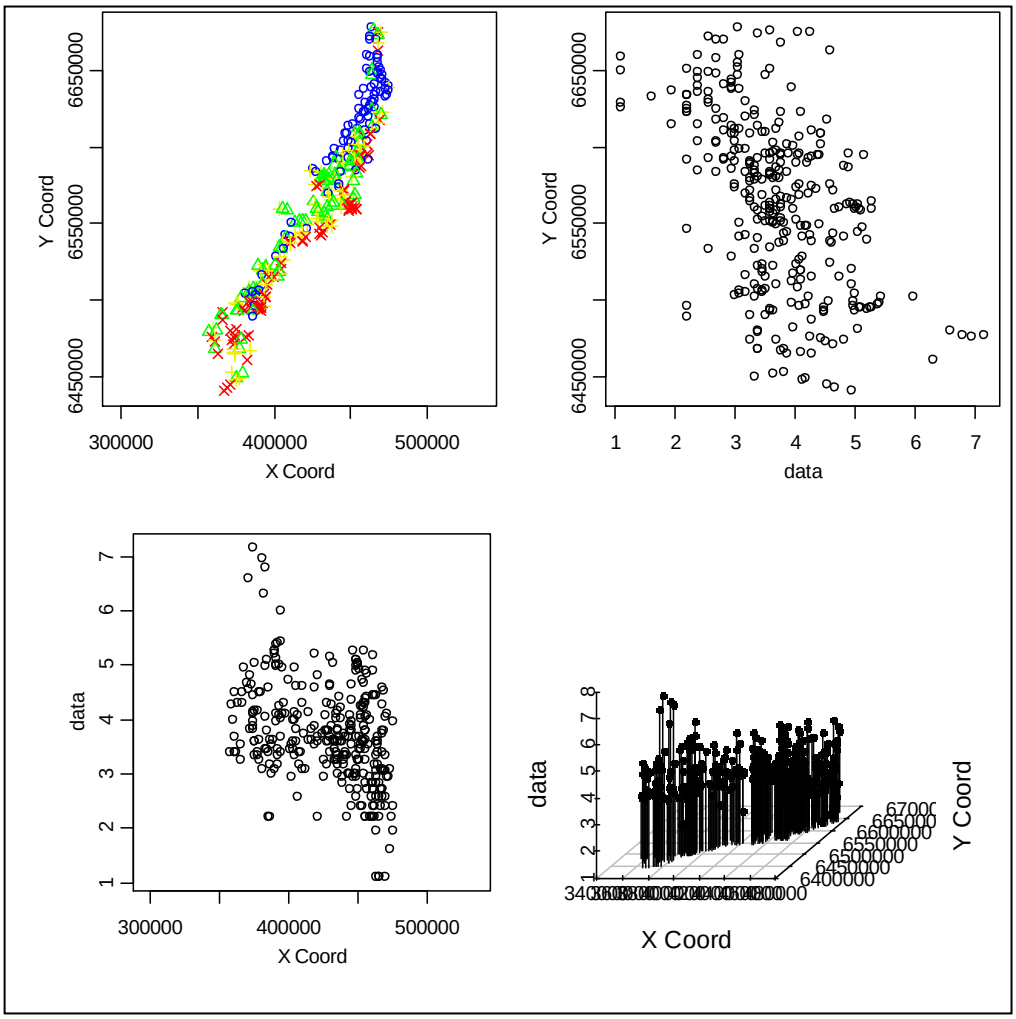

Figura 21 - Pontos localizados em suas coordenadas (topo esquerdo), valores de LogNa versus as coordenadas (topo direito e painel inferior esquerdo) e pontos dispersos sobre as coordenadas (painel inferior direito).

Os variogramas *cloud* para um modelo com e sem a inclusão da covariável, Figura 22, muda em relação à dispersão dos pontos na nuvem, no caso da inclusão ocorre uma diminuição na dispersão dos pontos. O variograma *bin* não muda quanto ao valor do efeito pepita e *sill*, embora a forma dos pontos é levemente afetada com a covariável. No *box-plots* para os *lags* não identifica-se grande modificação. Os valores para a variância dos dados (var.mark) e a máxima distância entre os pares de observações (max.dist), descritos na Tabela 12, não se alteram, mudando apenas as estimativa para os parâmetros da média do modelo (beta.ols) que depende da adição de uma tendência ao modelo.

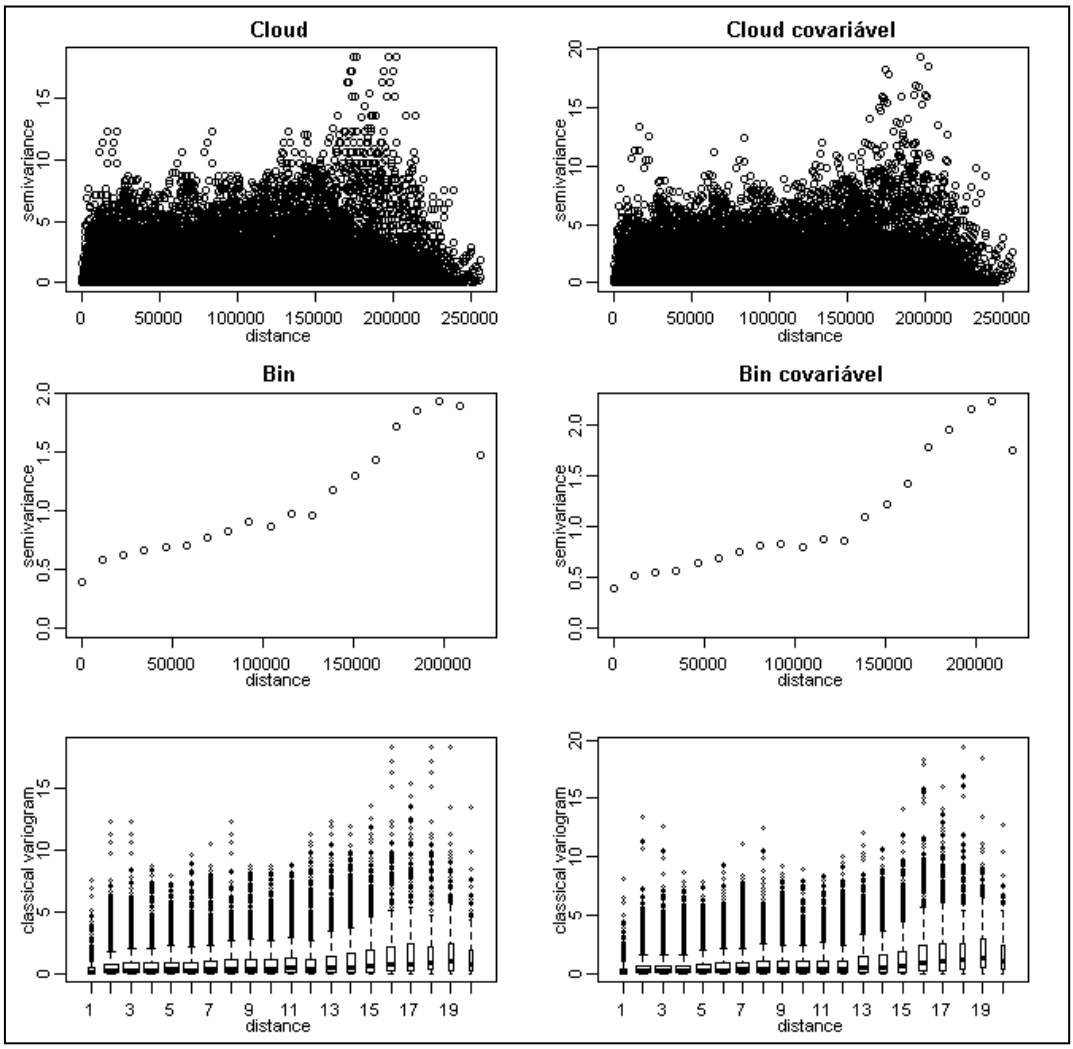

Figura 22 - Variograma cloud (superior), variograma bin (meio) e box-plot para os lags (inferior), sem a covariável (esquerda) e com a covariável (direita).

Tabela 12 - Resultados numéricos para os variogramas sem e com a covariável

LogNa	**Bin**	**Omnidirectional**		
tendência constante (s/. cov)	var.mark	beta.ols		max.dist
	0.860665	3.71082		225789.5
tendência distância (c/. cov)	var.mark	beta.ols		max.dist
	0.860665	3.994893	-0.00003329	225789.5

O ajuste para o variograma *bin*, assim como para a variável LogPST, será feita pelos quatro métodos de estimação diferentes - OLS, WLS, ML e REML. O modelo Gaussiano será empregado para expressar a dependência espacial dos dados. Este modelo apresenta a forma mais parecida ao variograma empírico por apresentar uma grande continuidade espacial. Para os ajustes especificou-se o mesmo efeito pepita para o variograma sem a covariável, não sendo o mesmo para quando se atribuiu a covariável ao modelo. Os valores iniciais exigidos pelos métodos ML e REML para o ajuste, foram especificados a partir do comando *eyefit* do pacote *geo*R, mudando quando tinha ou não a covariável no modelo. Os métodos OLS e WLS não exigem essa informação, embora os mesmos valores tenham sido especificados com o objetivo de manter um padrão na modelagem.

As retas ajustadas adicionadas ao variograma são mostradas na Figura 23, onde aparentemente os resultados para métodos ML e REML não resultam em ajustes satisfatórios. Estes métodos consideram todos os pontos do variograma *cloud*, o que pode causar uma impressão equivocada com respeito ao resultado do ajuste. Pelos resultados numéricos gerados pelos ajustes, verificou-se que a covariável é significativa para o método REML, cujo intervalo com 95% de confiança não contém o zero (Tabela 14). Existem evidências que apontam a covariável como sendo estatisticamente significativa. Para encontrar os valores apresentados nas Tabelas 14 e 15 realizou-se cada um dos ajustes para o modelo Gaussiano.

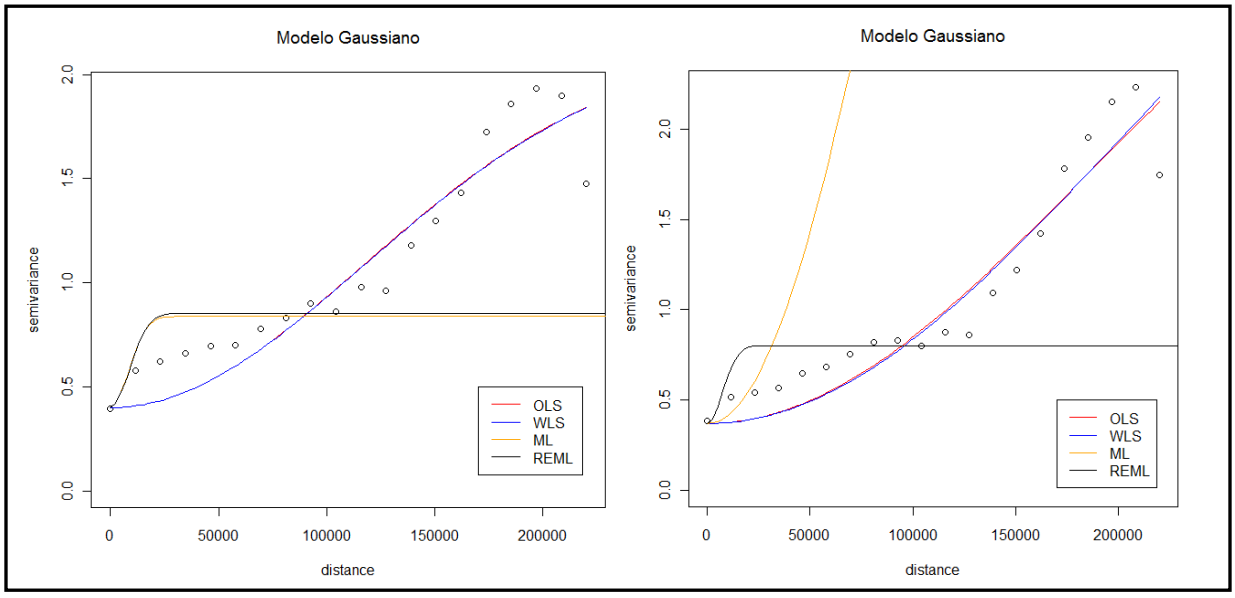

Figura 23 - Diferentes métodos de ajuste para o modelo Gaussiano sem a covariável (à esquerda) e com a covariável (à direita).

Tabela 13 – Resultados numéricos para o ajuste sem a covariável.

Método de ajuste	OLS	WLS	ML	REML
nugget effect	0.4	0.4	0.4	0.4
partial sill	1.736426	1.739276	0.4383	0.4546
range parameter	165268.4	165843.4	11739	12241
practical range	286049.6	287044.8	20317.57	21186.4
sum of squares	0.5016775	1018.030	---	---
maximised log-likelihood	---	---	-349.3	-347.4
AIC	---	---	704.6	700.8
BIC	---	---	715.9	712

Tabela 14 - Resultados numéricos para o ajuste com a covariável.

Método de ajuste	OLS	WLS	ML	REML
nugget	0.37	0.37	0.37	0.37
partial sill	3.51277	4.03092	12.3	0.4308
range parameter	261442.4	285108.1	166486	10370
practical range	452509.3	493470.3	288157.8	17948.35
sum of squares	0.62134	1063.216	---	---
maximised log-likelihood	---	---	-371.1	-344.1
covariável	---	---	-0.000020453	-0.000029362
IC covariável	---	---	(-0.000043389; 0.0000024833)	(-0.00005768; -0.0000010403)
AIC	---	---	750.2	696.1
BIC	---	---	765.2	711.2

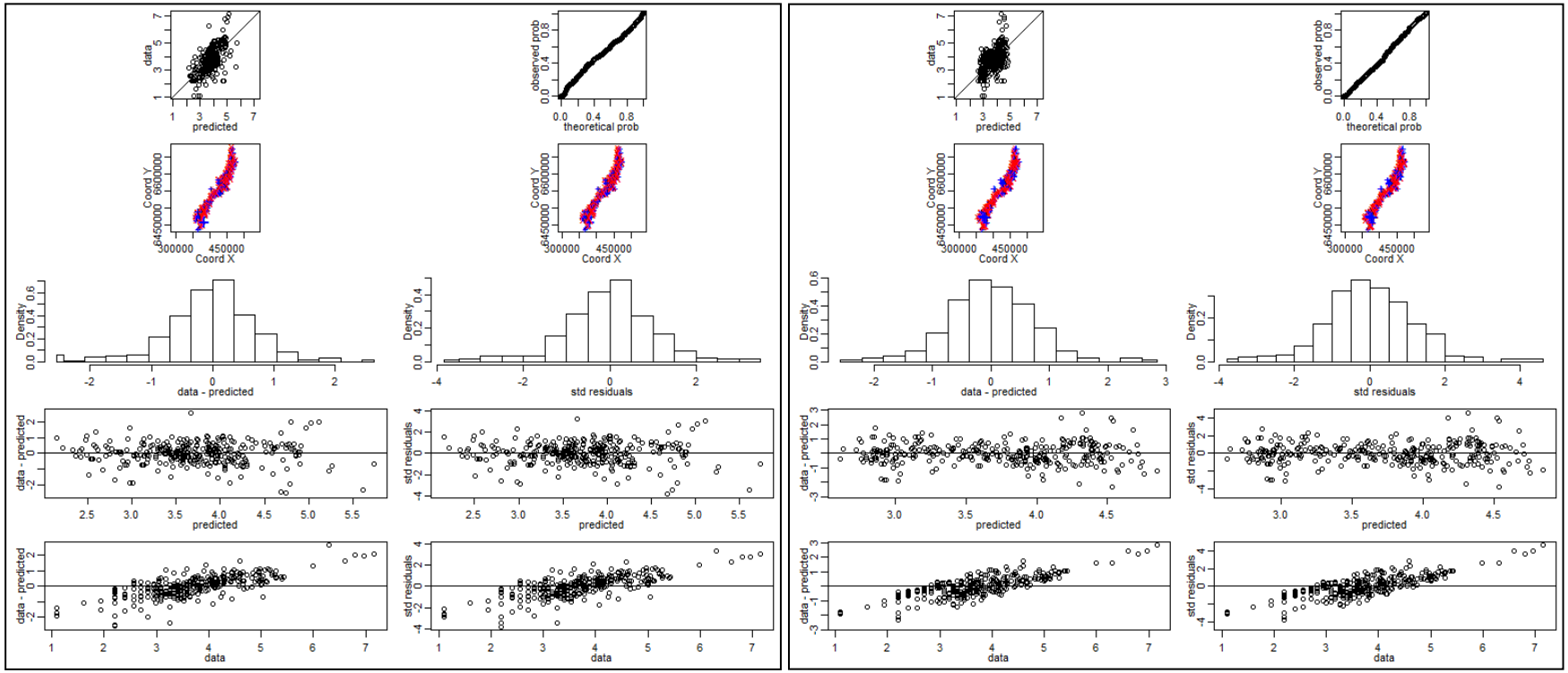

Figura 24 - Validação cruzada com covariável: método REML (à esquerda) e método ML (à direita).

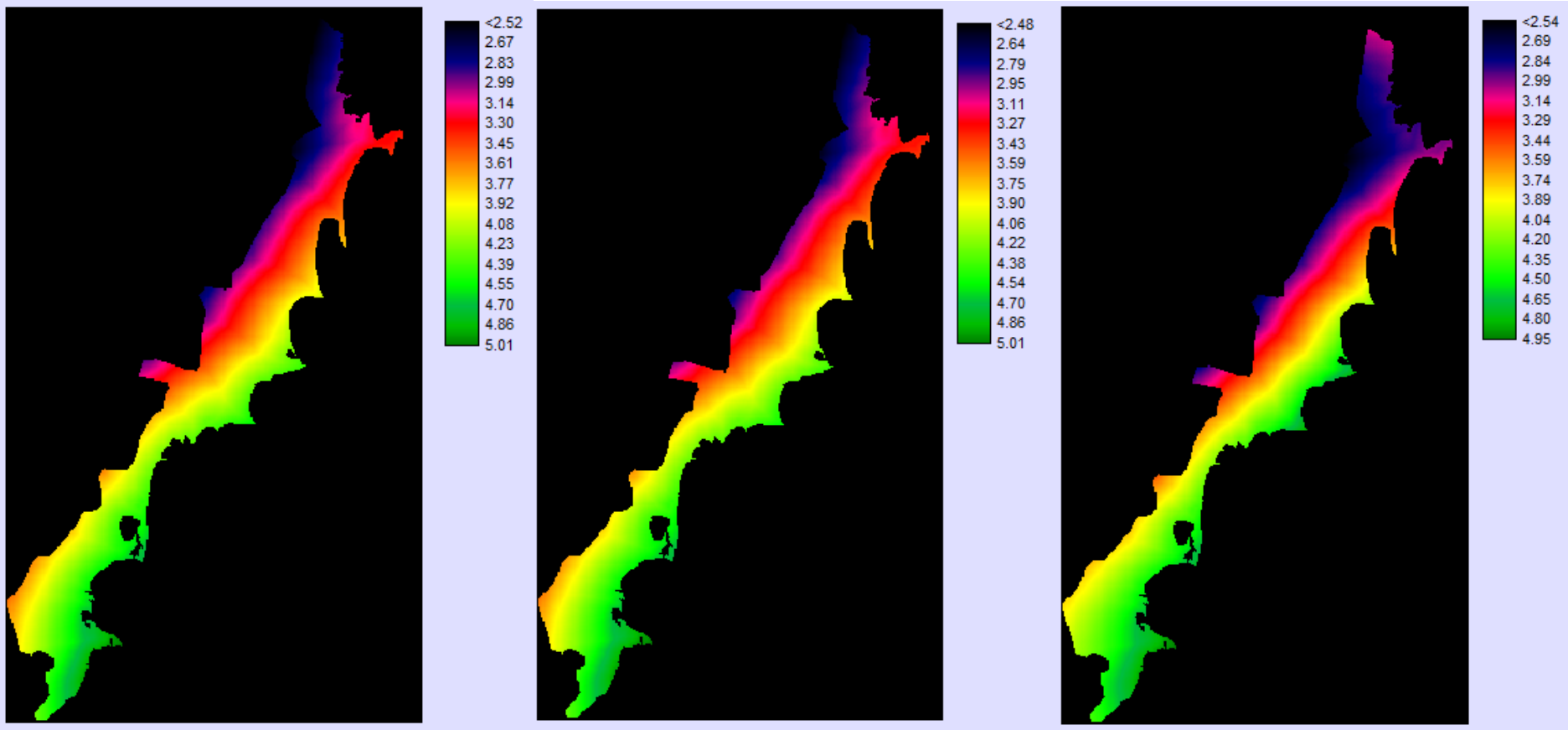

Figura 25 - Superfícies contínuas geradas pelos ajustes OLS, WLS e ML, respectivamente, com a covariável para o modelo Gaussiano.

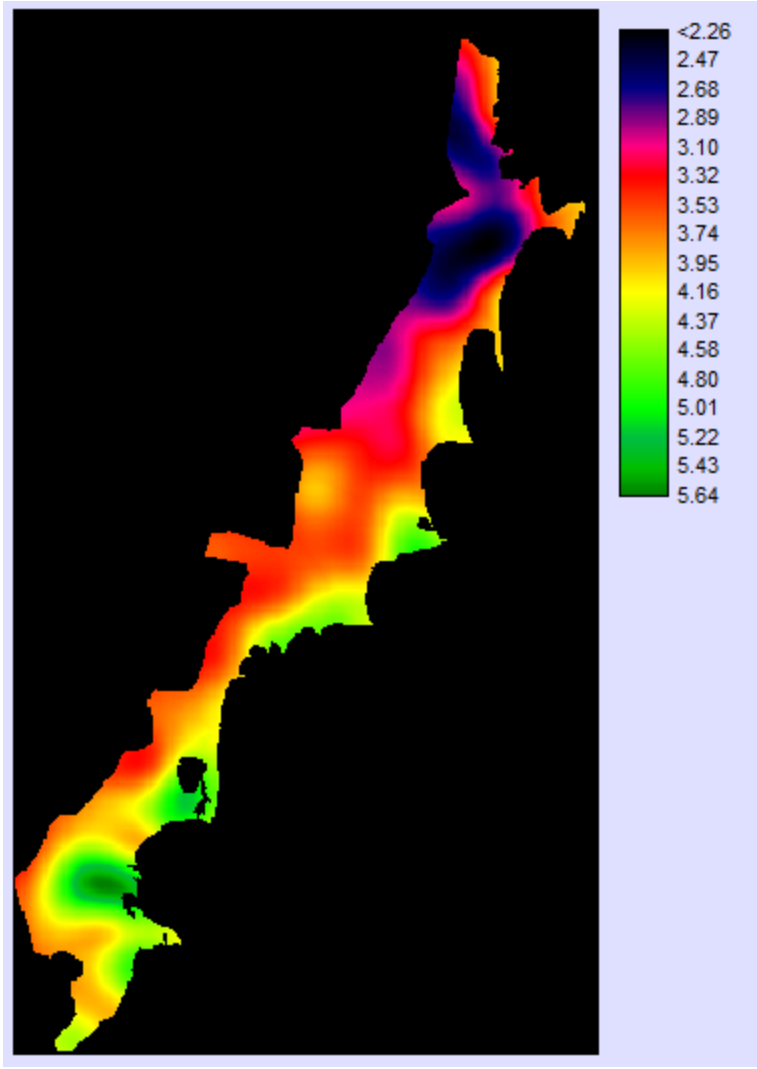

Figura 26 - Superfície contínua gerada pelo ajuste REML para o modelo Gaussiano com a covariável.

A Figura 24 apresenta o resultado gráfico da validação cruzada para os métodos ML e REML, nota-se uma amplitude menor dos valores reais menos os preditos pelo REML, indicando ser o método de estimação mais preciso para estimar os dados retirados da amostra em relação ao ML.

Com o objetivo de comparar os resultados gráficos, mapas de superfícies foram gerados a partir dos valores das estimativas para os parâmetros do modelo Gaussiano através dos métodos OLS, WLS e ML considerando a covariável, Figura 25. Poucas diferenças entre os dois primeiros mapas à esquerda são percebidas, mas em relação ao último mapa a diferença torna-se visualmente mais evidente. Contudo, estes mapas foram apresentados visando a comparação com a superfície gerada pelo método REML que apresentou a covariável, distância da margem da Laguna, como sendo significativa (Figura 26). A superfície resultante é bem diferente das geradas pelos métodos em que a significância não foi verificada, a amplitude da distribuição da salinidade vista pela escala de cores é maior, parece captar mais detalhes. Os resultados numéricos da validação cruzada descritos na Tabela 15, mostram que o método REML possui a menor amplitude entre os valores de máximo e mínimo para os erros calculados a partir da diferença entre os dados menos os valores preditos pela validação, a distância entre a média e a mediana é a menor e o desvio é menor.

Tabela 15 - Resultados numéricos da validação cruzada para os métodos de ajustes.

Método de ajuste	Min.	Mediana	Média	Máx.	Desvio
OLS c/ cov	-2.382	-0.06107	0.0011150	2.848	0.7511654
WLS c/ cov	-2.387	-0.05645	0.0011090	2.844	0.7524026
ML c/ cov	-2.339	-0.02098	0.0009072	2.839	0.736407
REML c/ cov	-2.541	0.01377	0.0007317	2.634	0.6962846

Visando avaliar a superfície, os dados observados da variável LogNa foram colocados sobre ela para averiguar o desempenho do ajuste feito pelo método REML para o modelo Gaussiano, Figura 27. A superfície gerada pela krigagem Ordinária apresenta a distribuição do fenômeno na região de estudo caracterizada pelos valores observados da variável LogNa.

Como resultado, o método REML foi escolhido para ajustar o modelo Gaussiano a fim de descrever a estrutura de covariância dos dados, e suas estimativas (Tabela 15 coluna 50) foram utilizadas para predizer valores nos locais não observados na região.

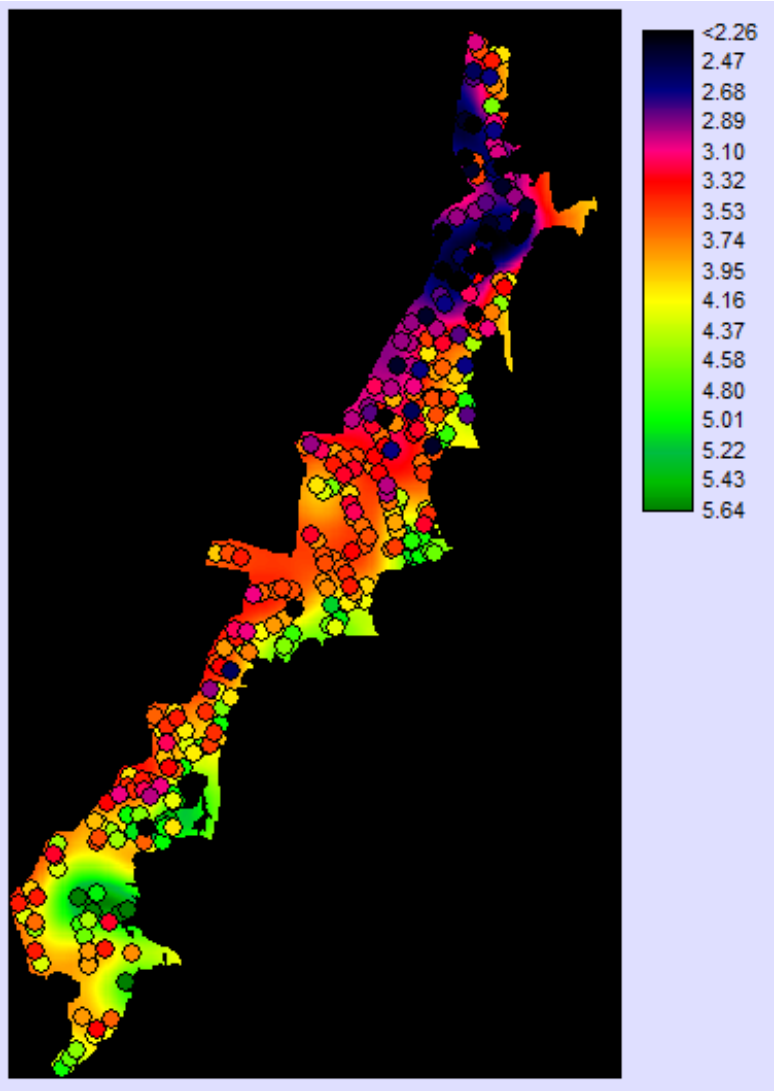

Figura 27 - Sobreposição dos dados na superfície gerada pelo método REML.

3. 3. RESULTADOS VARIÁVEL CE

A análise da variável condutividade elétrica (CE) seguirá o mesmo desenvolvimento apresentado para as variáveis PST e Na. Procedendo a análise exploratória, com as medidas de posição apresentadas na Tabela 16, verifica-se que a amplitude entre o valor mínimo e máximo nos dados é grande, no entanto a média e a mediana não possuem valores afastados. Para identificar assimetria nos dados, verificada nas variáveis anteriores, gerou-se o histograma (Figura 28). Das 315 observações medidas para a variável CE no BLOCO 1, 310 são mais frequentes entre os valores 0 e 5, evidenciando uma distribuição muito assimétrica à esquerda. De forma análoga ao procedimento adotado antes, os dados foram transformados para ln (CE+1), e identificados agora por LogCE. O resultado para as medidas de posição do LogCE, Tabela 17, mostram uma redução na amplitude entre o valor mínimo e máximo bem como na distância entre a média e mediana; o histograma para a variável transformada mostra os dados mais distribuídos, porém a assimetria à esquerda continua existindo. O *box-plot* para os dois casos (Figura 29), dados originais e transformados, indica que apesar da transformação ter aproximado mais os valores, dados atípicos continuam afetando a distribuição dos dados.

Tabela 16 - Medidas de posição para os dados da variável CE.

CE			
Min.	Mediana	Média	Max.
0.300	1.000	1.339	25.000

Tabela 17 - Medidas de posição para os dados da variável LogCE.

LogCE			
Min.	Mediana	Média	Max.
0.2624	0.6931	0.7590	3.2580

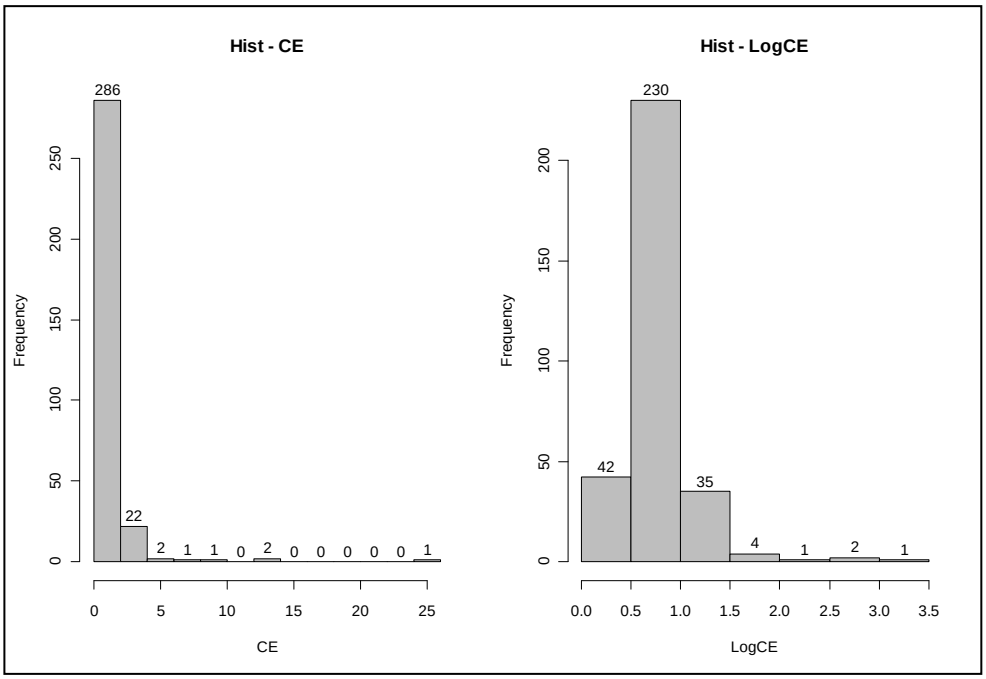

Figura 28 - Histograma para os dados originais (à esquerda) e os dados transformados (à direita).

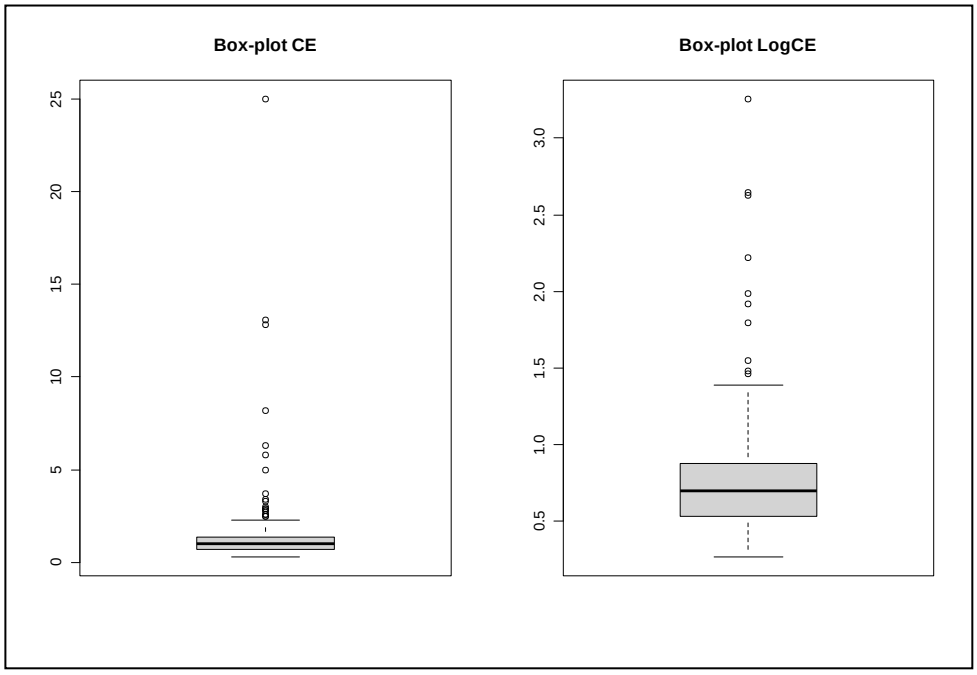

Figura 29 – Box-plot para os dados originais (à esquerda) e os dados transformados (à direita).

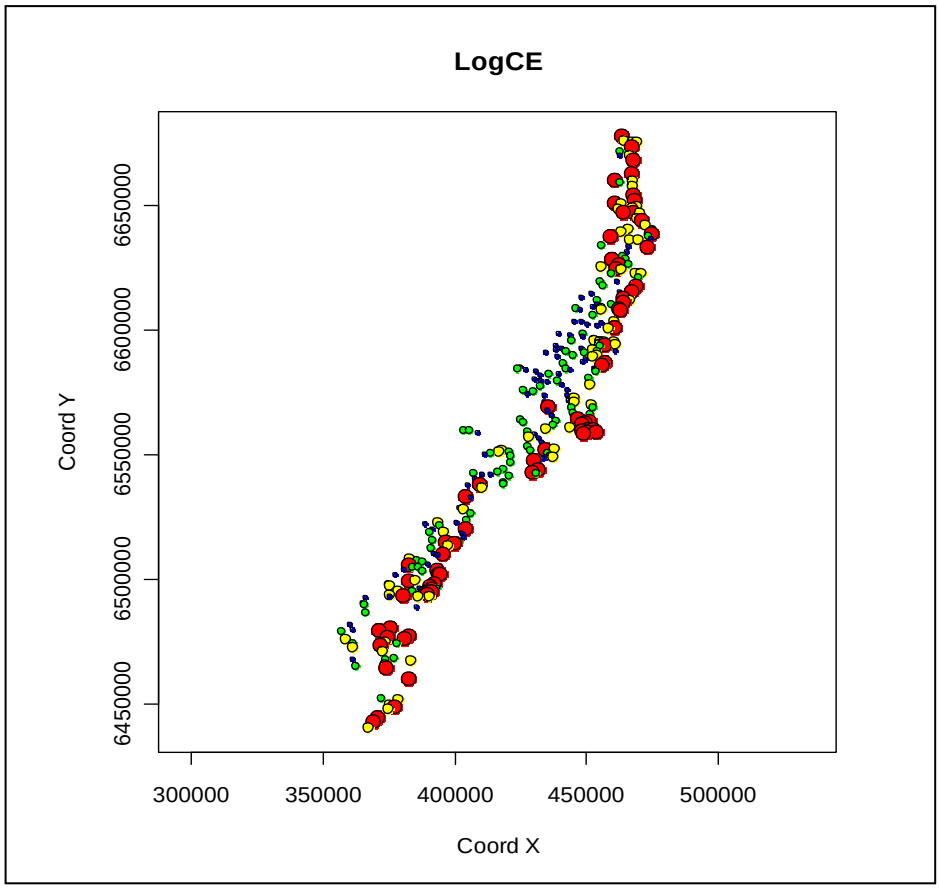

Figura 30 - Gráfico mostrando a localização dos 315 pontos amostrados para o LogCE.

A Figura 30 representa a distribuição da variável transformada LogCE localizadas em suas respectivas coordenadas geográficas. Diferentemente das outras variáveis, a condutividade elétrica (CE) apresenta valores altos para os pontos localizados em toda a margem da Laguna. Possivelmente a covariável distância da margem da Laguna será importante na modelagem desta variável.

Analisando a Figura 31, constata-se uma concentração da dispersão dos dados da variável LogCE tanto em relação a coordenada Y como na coordenada X, embora um pouco mais dispersos em Y. Os pontos dispersos sobre as coordenadas (painel inferior direito) parecem estar distribuídos de maneira uniforme no centro, valores atípicos em relação a vizinhança são verificados nas extremidades, visto que a altura de cada ponto corresponde ao valor medido naquele ponto.

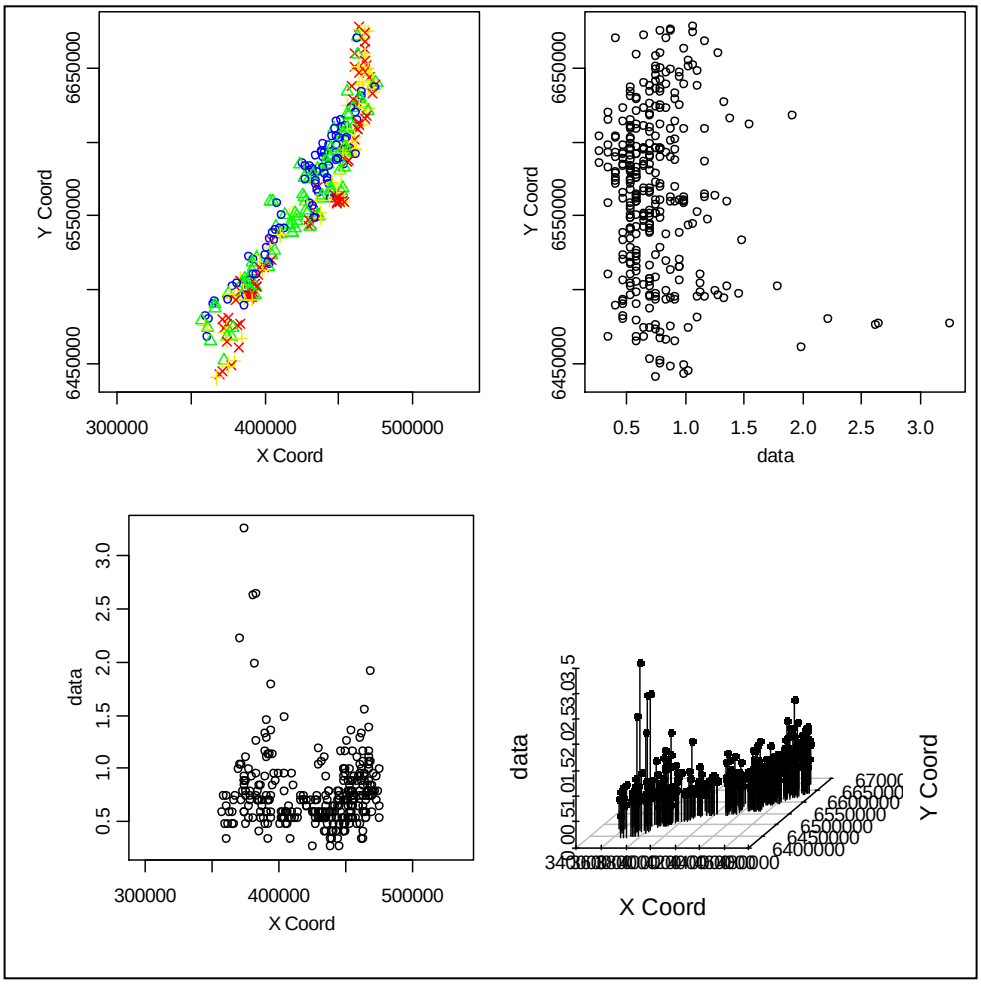

Figura 31 - Pontos localizados em suas coordenadas (topo esquerdo), valores de LogCE versus as coordenadas (topo direito e painel inferior esquerdo) e pontos dispersos sobre as coordenadas (painel inferior direito).

Os variogramas *cloud* (Figura 32) para o modelo com e sem a covariável apresentam uma grande dispersão de pontos da nuvem considerando a distância dos pares individualmente. O variograma *bin* parece ter uma flutuação menor dos pontos na medida em que a distância aumenta, o valor do efeito pepita e do patamar não mudaram. O *box-plots* para os *lags* não apresentaram visualmente grande modificação.

Analisando os resultados numéricos descritos na Tabela 18, os valores para a variância dos dados (var.mark) e a máxima distância entre os pares de observações (max.dist) não se alteram. Já, os parâmetros da média do modelo (beta.ols) dependem da quantidade de covariáveis assumidas no modelo.

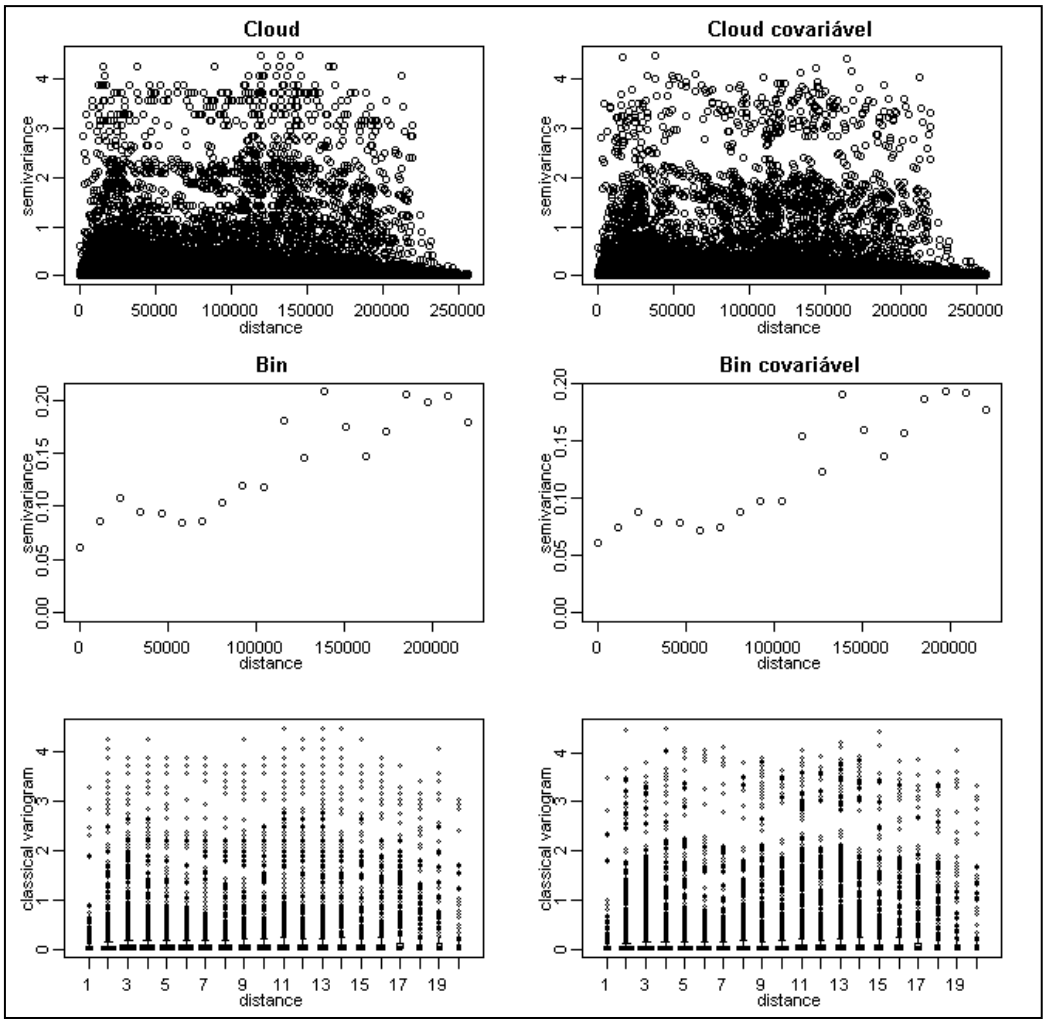

Figura 32 - Variograma cloud (superior), variograma bin (meio) e box-plot para os lags (inferior), sem a covariável (esquerda) e com a covariável (direita).

Tabela 18 - Resultados numéricos para os variogramas sem e com a covariável

LogCE	**Bin**	**Omnidirectional**		
tendência constante (s/. cov)	var.mark	beta.ols		max.dist
	0.1174352	0.7590052		225789.5
tendência distância (c/. cov)	var.mark	beta.ols		max.dist
	0.1174352	0.92827	-0.00001984	225789.5

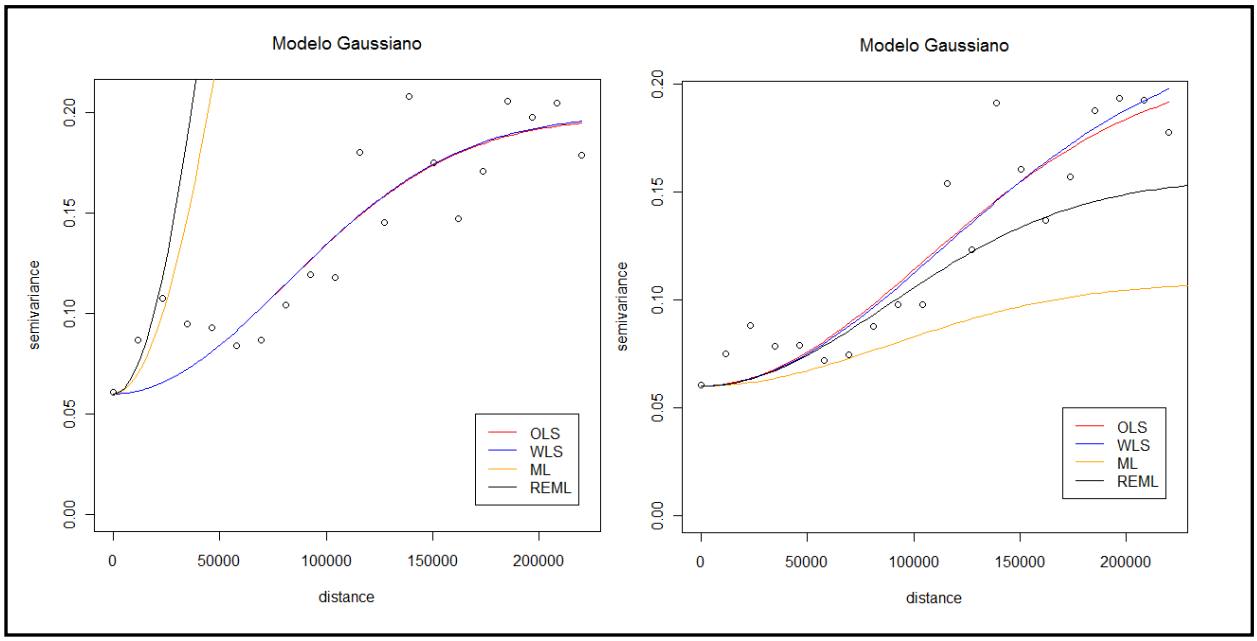

Figura 33 - Diferentes métodos de ajuste para o modelo Gaussiano sem a covariável(à esquerda) e com a covariável (à direita).

Para ajustar uma curva ao variograma empírico (variograma *bin*) empregou-se os quatro métodos de estimação, OLS, WLS, ML e REML, e o modelo Gaussiano para expressar a dependência espacial dos dados. Para os ajustes especificou-se o mesmo efeito pepita para o variograma com e sem a covariável, os valores iniciais exigidos pelos métodos ML e REML para o ajuste foram especificados a partir do comando *eyefit* do pacote *geo*R, mudando os valores quando o modelo incluía a covariável. Como citado anteriormente, os métodos OLS e WLS não exigem essa informação, mas mantiveram-se os mesmos valores para seguir um "padrão" na modelagem.

As retas ajustadas ao variograma são apresentadas na Figura 33. Sem considerar a covariável, os métodos OLS e WLS parecem estar sobrepostos e se ajustam bem aos pontos, enquanto que novamente os métodos ML e REML não parecem resultar um bom ajuste. Como estes métodos consideram todos os pontos do variograma *cloud*, o resultado visual não pode ser o único meio de decisão. As curvas para os ajustes considerando a covariável evidenciam uma diferença em relação às curvas sem a covariável, principalmente para os métodos ML e REML que visualmente se aproximaram mais dos pontos do variograma.

As Tabelas 19 e 20 apresentam os resultados gerados pelos ajustes sem considerar a covariável no modelo e considerando-a no modelo, respectivamente. Verifica-se que a distância da margem da Laguna (covariável) é estatisticamente significativa para o método ML e REML já

que intervalo com 95% de confiança não contém o zero absoluto. Desta forma, existem evidências amostrais indicando que a covariável é significativa para o modelo.

Como não foi possível gerar um intervalo para testar a significância da covariável para os métodos OLS e WLS, a validação cruzada será feita para as estimativas resultantes destes métodos, assim como para ML e REML visando a comparação dos resultados. O objetivo é verificar a precisão do modelo que expressa a estrutura de covariância espacial dos dados a partir dos métodos. Avaliando visualmente a validação para cada um dos métodos (Figuras 34 e 35), não se identifica diferença significativa. A Tabela 21 descreve os resultados numéricos para a validação cruzada para os quatro diferentes métodos de estimação, todos apresentam valores próximos.

Os mapas de superfícies gerados pelas estimativas dos parâmetros são descritos nas Figuras 36 e 37, correspondem respectivamente aos métodos OLS e WLS, e ML e REML. A amplitude de predição, representada pela escala de cores, na descrição da distribuição da variável LogCE sobre toda a região de estudo apresentaram pouca diferença. A distribuição da salinidade representada através da superfície quando comparadas entre os métodos de ajustes não mostraram mudanças relevantes sobre o mapa.

Usando a informação da significância da covariável no modelo para os métodos ML e REML, avaliou-se o desempenho dos dois através da sobreposição dos dados da variável LogCE na superfície estimada (Figura 38). A análise visual do método de ajuste REML parece refletir melhor o comportamento dos dados na região em algumas partes (círculo na Figura 38). Para auxiliar a visualização, foi feito um *zoom* nos mapas gerados por estes métodos (Figura 39) a fim de identificar e avaliar o desempenho deles. Identificaram-se poucas diferenças (apontadas por círculos na figura) porém as encontradas mostram que o método REML capta um pouco mais o comportamento da salinidade representados pelos dados; apesar dos resultados obtidos pelo método ML serem parecido com o REML.

O método de estimação da Máxima Verossimilhança Restrita para o modelo Gaussiano é adequado para expressar a salinidade na região estudada através das medidas da variável LogCE tanto para descrever a variabilidade espacial quanto para predição.

Tabela 19 - Resultados numéricos para o ajuste sem a covariável.

Método de ajuste	OLS	WLS	ML	REML
nugget effect	0.06	0.06	0.06	0.06
partial sill	0.13807	0.139083	1.076	1.567
range parameter	114040.7	114435.2	118919	118919
practical range	197383.8	198066.5	205827.1	205827.1
sum of squares	0.008787848	26.16765	---	---
maximised log-likelihood	---	---	-98.11	-94.64
AIC	---	---	202.2	195.3
BIC	---	---	213.5	206.5

Tabela 20 - Resultados numéricos para o ajuste sem a covariável.

Método de ajuste	OLS	WLS	ML	REML
nugget	0.06	0.06	0.06	0.06
partial sill	0.147982	0.162419	0.0482	0.0962
range parameter	148670.4	160518.2	124865	124865
practical range	257321.4	277827.8	216118.4	216118.4
sum of squares	0.006092654	14.7492	---	---
maximised log-likelihood	---	---	-91.89	-89.96
covariável	---	---	-0.00001849	-0.00001759
IC covariável	---	---	(-0.00002406; -0.000012917)	(-0.00002395; -0.000011238)
AIC	---	---	191.8	187.9
BIC	---	---	206.8	202.9

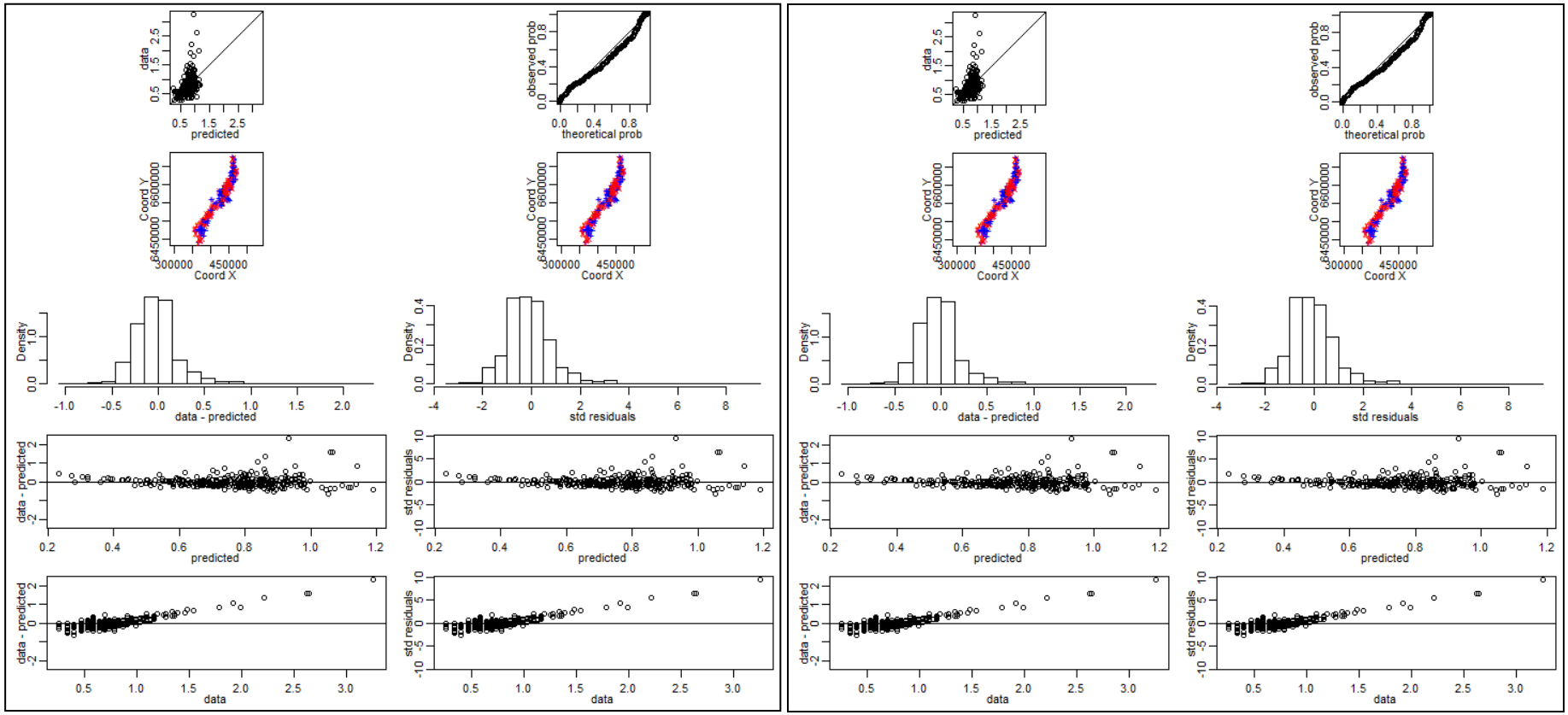

Figura 34 – Validação cruzada para os métodos de ajuste OLS e WLS com a covariável.

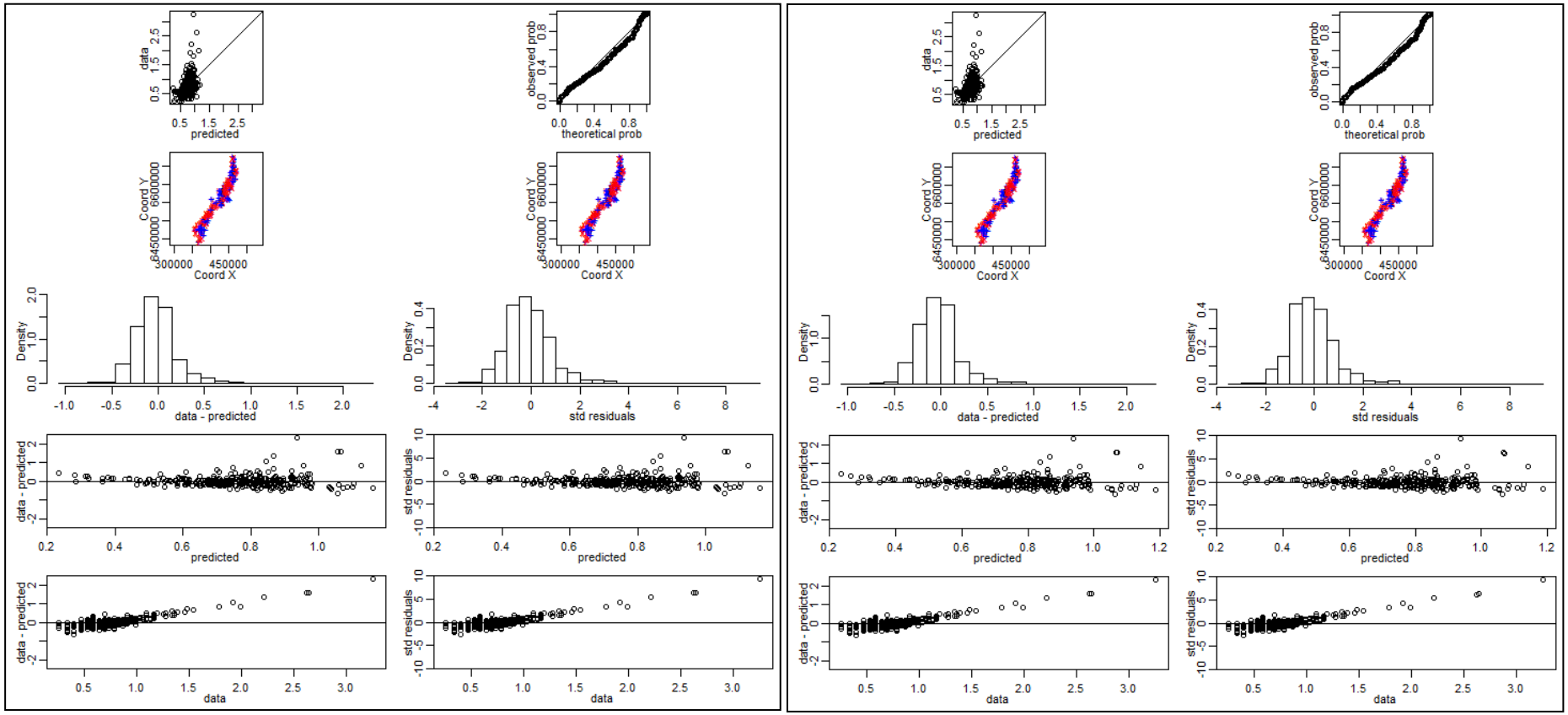

Figura 35 - Validação cruzada para os métodos de ajuste ML e REML com a covariável.

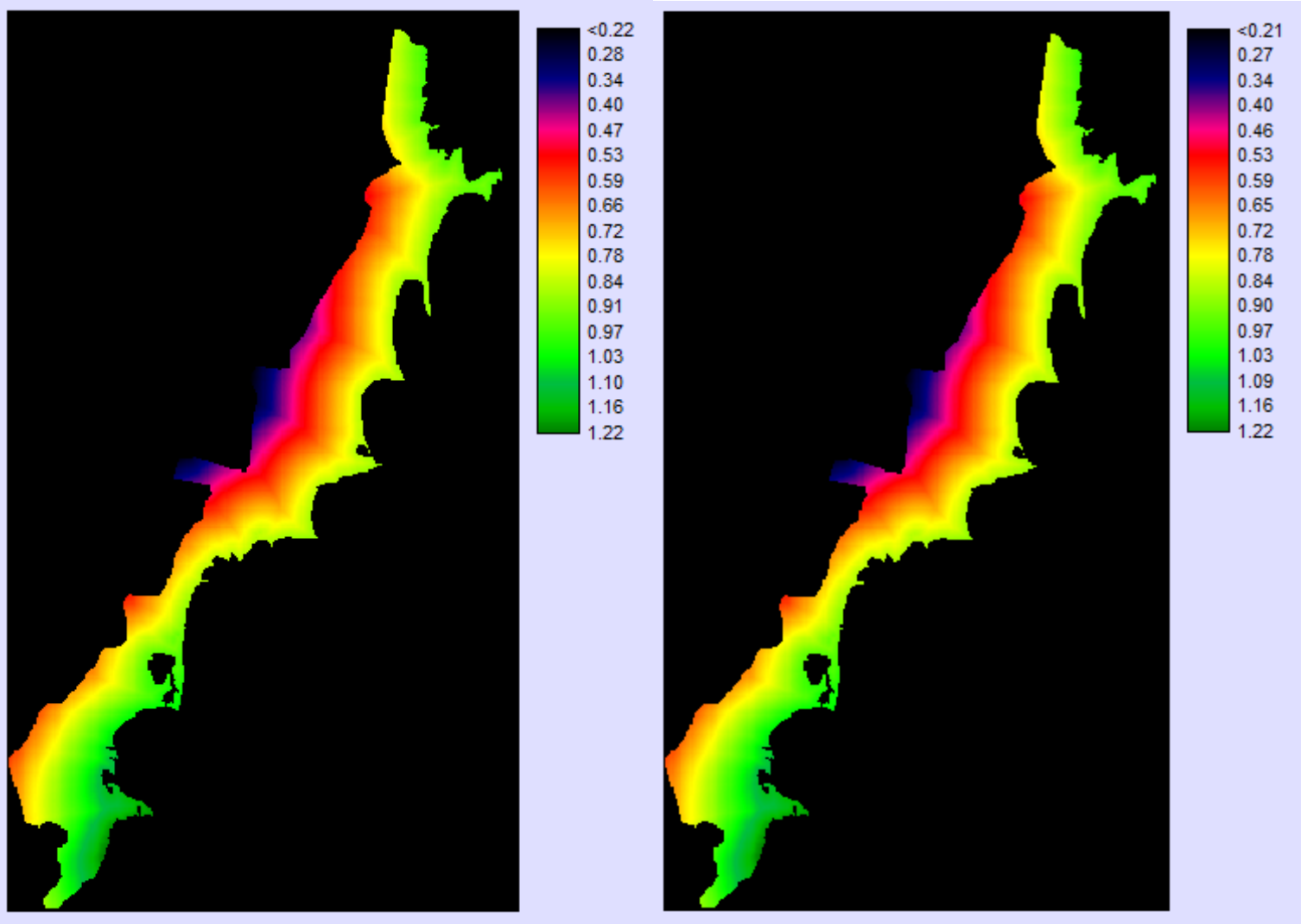

Figura 36 – Superfícies contínuas geradas pelas estimativas dos ajustes OLS e WLS para o modelo Gaussiano considerando a covariável.

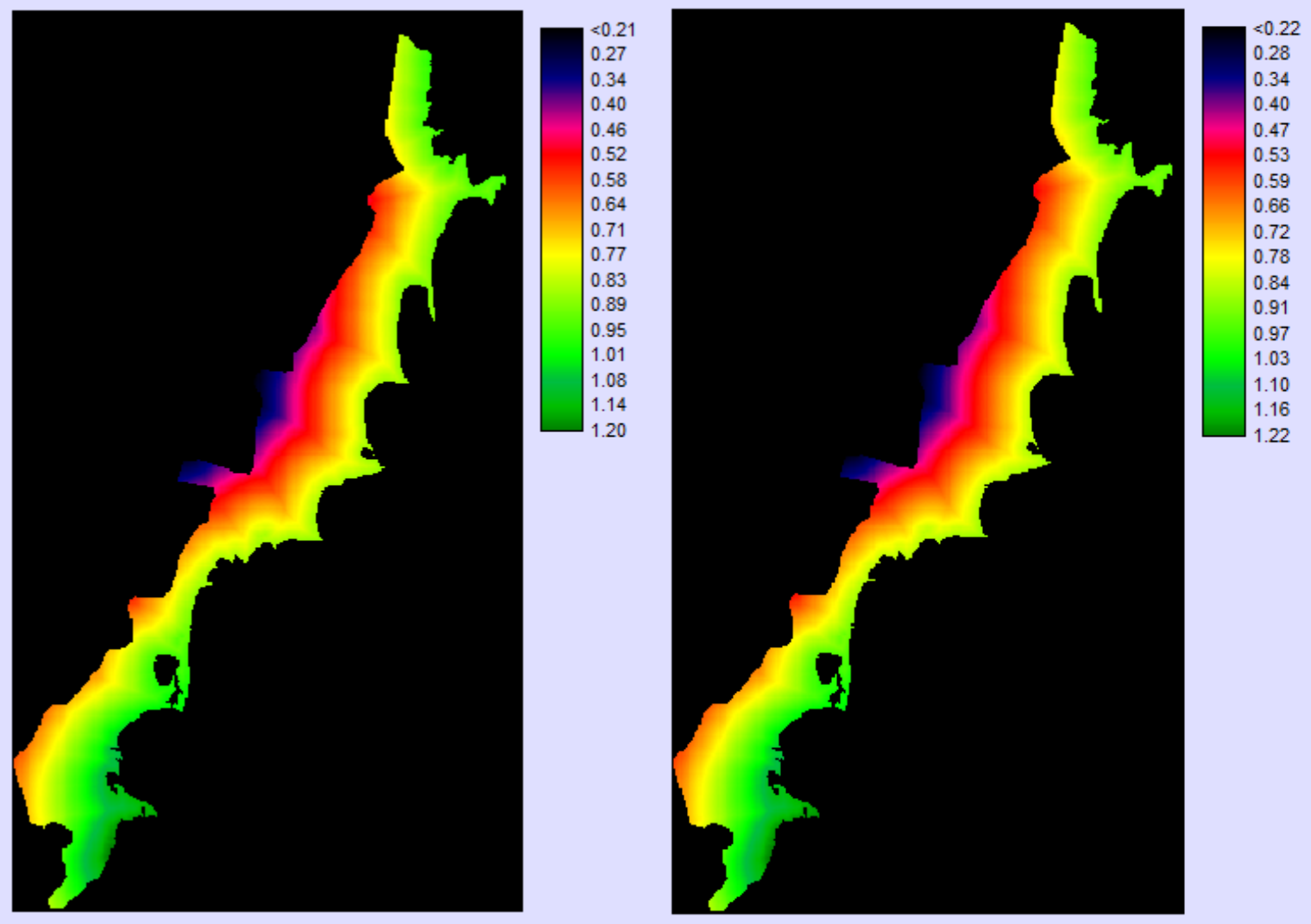

Figura 37 - Superfícies contínuas geradas pelas estimativas dos ajustes ML e REML para o modelo Gaussiano considerando a covariável.

Tabela 21 - Resultados numéricos da validação cruzada para os métodos de ajustes.

Método de ajuste	Min.	Mediana	Média	Máx.	Desvio
OLS c/ cov	-0.6468	-0.04215	0.000002058	2.325	0.305952
WLS c/ cov	-0.6425	-0.04239	-0.00000769	2.327	0.3060508
ML c/ cov	-0.6497	-0.04228	0.00008428	2.323	0.3059298
REML c/ cov	-0.6564	-0.04349	0.00004294	2.320	0.3057820

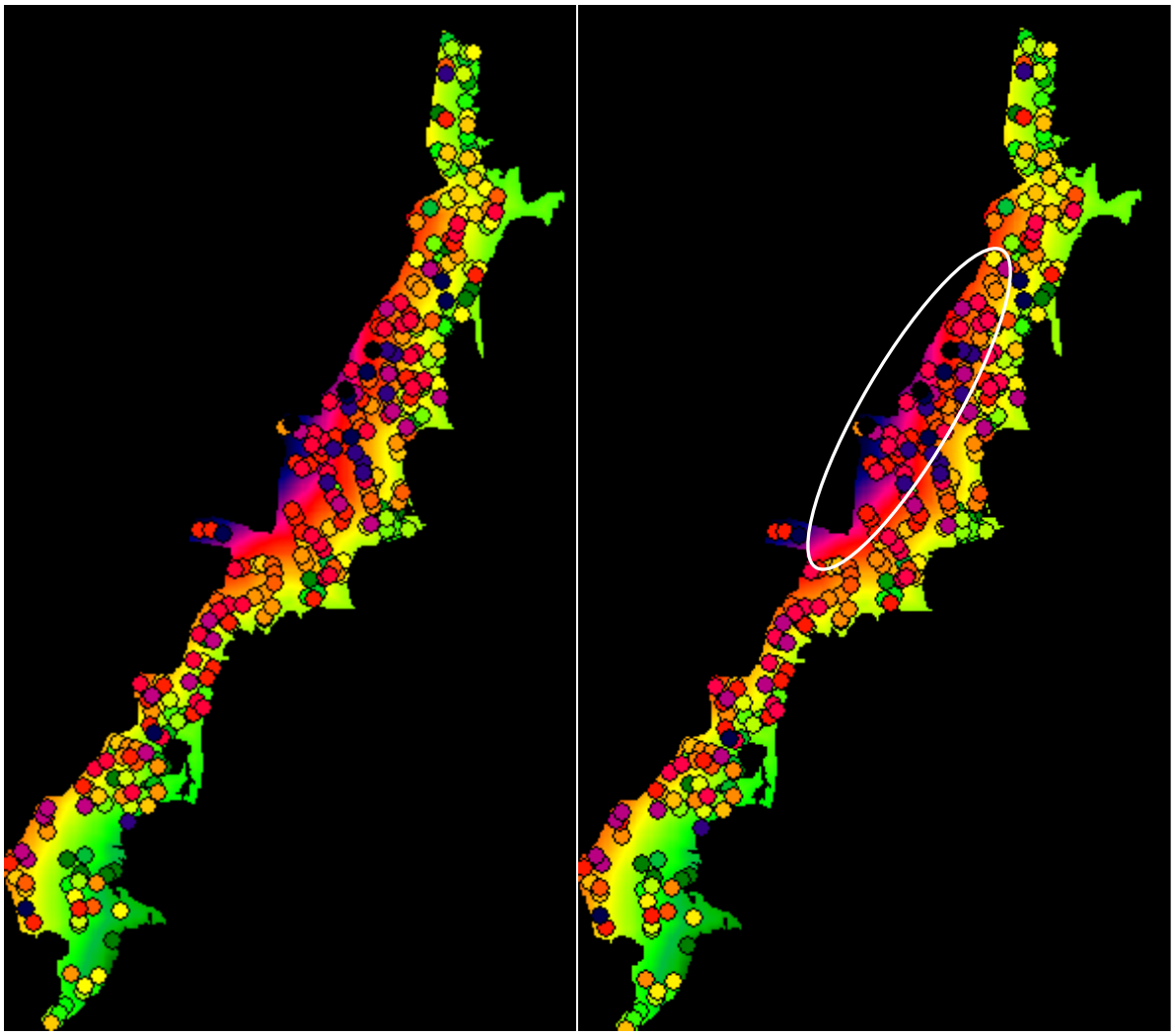

Figura 38 – Sobreposição dos dados na superfície gerada pelos métodos de ajuste ML (à esquerda) e REML (à direita).

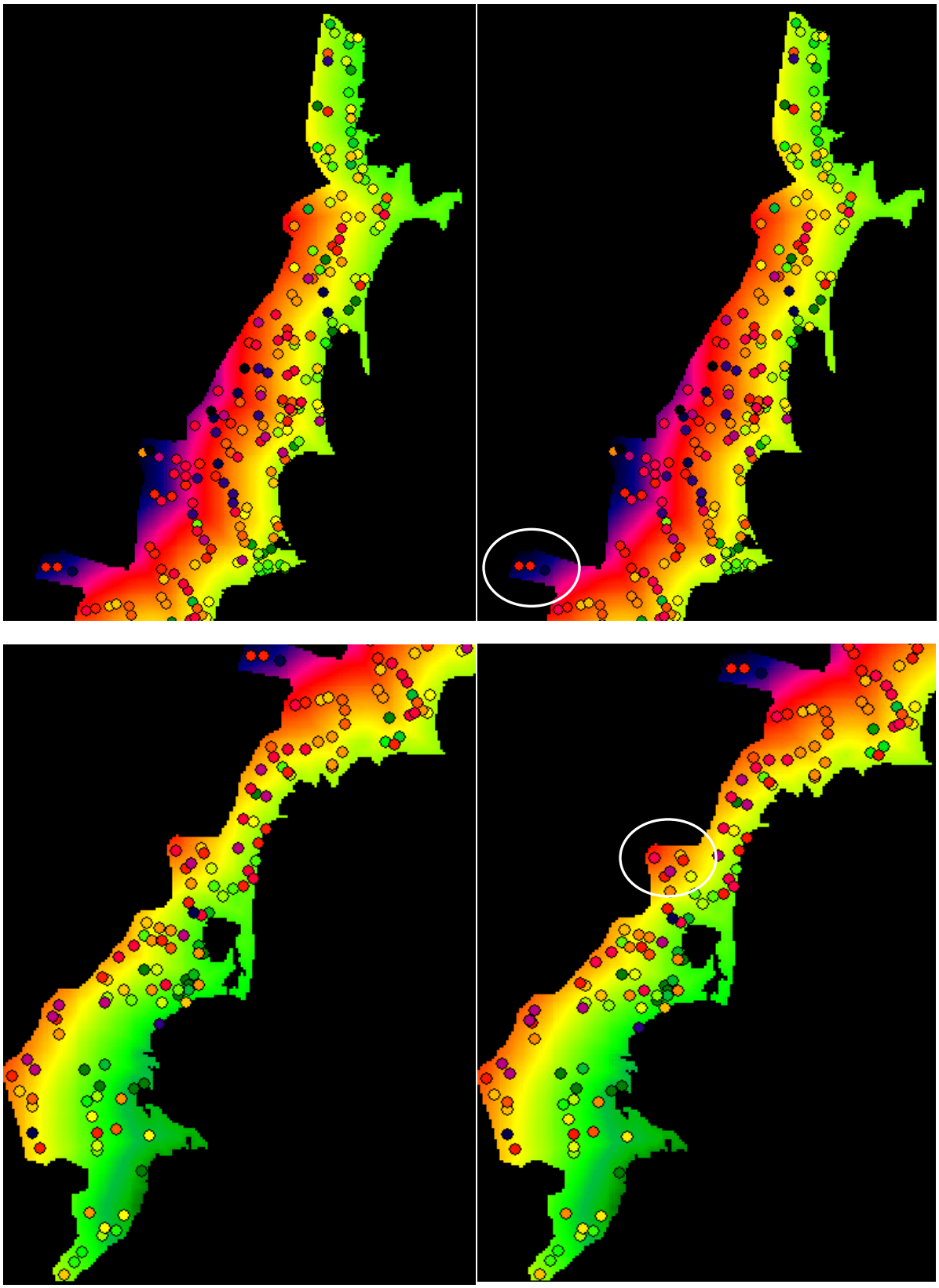

Figura 39 – Zoom nas superfícies geradas pelos métodos de ajuste ML (à esquerda) e REML (à direita).

4. DISCUSSÃO

As três variáveis analisadas apresentaram uma distribuição assimétrica causada principalmente pelos valores atípicos identificados numericamente pelas medidas de posição e graficamente pelo histograma e *box-plots*, sendo percebida uma influência maior na variável condutividade elétrica. Visando contornar esse problema realizou-se a transformação dos dados, a transformação escolhida é bastante usada nos dados decorrentes de medidas de variáveis químicas. Avaliando o resultado da transformação, verificou-se que o problema foi contornado pois ao comparar os dados originais com os transformados ficou nítida a mudança na distribuição dos dados, o que mostra que a escolha por essa transformação a todas as variáveis foi adequada.

Quando se executou a análise exploratória espacial dos dados graficamente, identificou-se uma possível tendência na região caracterizada pela Distância da margem da Laguna. Como estratégia de análise, decidiu-se analisar os dados com e sem a inclusão de uma covariável no modelo. Com o objetivo de avaliar as duas situações, todos os procedimentos desenvolvidos consideraram essas duas possibilidades. No entanto, quando a significância estatística se confirmava, o modelo que a incluía era comparado aos modelos que obtiveram bons resultados na validação cruzada. Porém nos casos em que se evidenciou a importância da covariável no modelo, os resultados se diferenciaram dos outros (diferença mais evidente na variável Na).

Os variogramas *cloud* e *bin*, que auxiliam na caracterização da estrutura de covariância, foram construídos para duas situações e mostram que os dados possuem grande continuidade espacial, o que levou a escolha do modelo Gaussiano como modelo teórico para descrever a variabilidade espacial para as três variáveis. Após a escolha do modelo, foi preciso especificar o método de ajuste cuja estratégia permitiu comparar quatro formas de estimação disponíveis no pacote *geo*R. Uma das informações que o programa possibilita especificar são os valores iniciais e o valor do efeito pepita para o processo, sendo o primeiro uma exigência do método da Máxima Verossimilhança e Máxima Verossimilhança Restrita, que, se não informado, não fornece resultado. Para cada uma das variáveis, os valores definidos foram mantidos os mesmos para os quatro métodos de ajuste mudando apenas quando a covariável era ou não considerada.

Os resultados mostram que neste caso os métodos dos Mínimos Quadrados foram parecidos entre si e diferentes dos métodos da Máxima Verossimilhança, tanto na comparação da validação cruzada quanto nas superfícies de predição geradas.

Na modelagem geoestatística decisões são tomadas a cada passo. Inicialmente a análise exploratória indica o caminho a ser seguido, com a transformação ou não dos dados; a construção do variograma depende da investigação do comportamento dos dados para diferentes direções o que permite identificar se o processo é isotrópico ou anisotrópico; a escolha do modelo mais adequado depende do comportamento dos pontos do variograma empírico; o método de ajuste que resultará nas estimativas dos parâmetros do modelo teórico e que serão usadas como informação a priori para o método de interpolação empregado para gerar uma superfície contínua que apresenta a distribuição espacial do fenômeno sobre a toda a região de estudo.

A escolha do modelo teórico e do método de ajuste pode variar dependendo dos valores especificados. Outros resultados podem ser gerados o que torna muito ampla e aberta a análise dos dados geoestatísticos sob enfoque clássico.

5. CONCLUSÃO

Os resultados do ajuste geralmente parecem ser satisfatórios, porém a questão que fica é: o quanto o variograma empírico e o teórico são ótimos e se ajustam bem?

Os critérios adotados na modelagem podem variar e isso muda o resultado final; a presença do pesquisador que conhece como o fenômeno ocorre na região seria uma alternativa para lidar com as inúmeras possibilidades de escolha que podem ser tomadas ao longo da análise.

Cada vez mais a geoestatística alcança diversas áreas de pesquisa o que leva muitos pesquisadores a modelar os dados sem verificar as suposições dos modelos geoestatísticos.

Neste estudo a modelagem clássica foi desenvolvida, os dados reais para cada variável referente à salinidade na região tornaram a análise difícil porque a instabilidade verificada afetava os resultados, e influenciava na compreensão do fenômeno.

A investigação da salinidade na porção oeste da Laguna dos Patos poderia seguir uma nova estratégia para modelar os dados, como a inclusão de uma segunda covariável: Distância da entrada do oceano na extremidade sul. Por causa da distribuição espacial observada para os valores das variáveis LogPST e LogNa considerando a localização geográfica, por apresentarem um padrão espacial caracterizado por esta distância.

REFERÊNCIAS BIBLIOGRÁFICAS

Adler, R. J. (1981). *The Geometry of Random Fields*. Wiley, New York.

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

Diggle, P. J. e Ribeiro Jr, P. J. (2000). *Model Based Goestatistics*. 14º SINAPE, Associação Brasileira de Estatítica – ABE.

Diggle, P. J., Ribeiro Jr, P. J. e Christensen, O. L. (2003). An Introduction to Model-based Geoestatistics. *Spatial statistics and computational methods, Cap.2,* J. Moller, ed. Springer.

Diggle, P. J. e Ribeiro Jr, P. J. (2007). *Model-based Geoestatistics*. Springer, New York.

Ferreira, G. S. (2002). *Geoestatística: Estimação e Predição supondo um Processo Gaussiano Subjacente*. Monografia para obtenção do Grau de Bacharel em Estatística, Instituto de Matemática, UFRGS.

Journel, A. G. e Huijbregts, C. J. (1978). *Minning Geoestatistics*. Academic Press, London.

Matheron, G. (1963). Principles of Geoestatistics. *Economic Geology*, 58, 1246-1266.

Patterson, H. D. e Thompson, R. (1974). *Maximum likelihood estimation of components of variance*. Proceedings of the 8th International Biometric Conference, Biometric Society, Washington, DC, 197-207.

Pulgati, F. H. (2004). *O uso de modelos bayesianos na definição de áreas espaciais alteradas por atividade de perfuração exploratória marítima*. Tese para obtenção do Grau de Mestre em Ciências em Engenharia Civil, UFRJ.

Rosenbaltt, M. (1985). *Stationary Sequences and Random Fields*. Birkhauser, Boston.

ANEXO

Rotinas do pacote *geo*R usados na análise dos dados da monografia.

```
require(geoR)
# CARREGANDO OS DADOS
dB <- read.geodata("banco.completo.txt", header=T, coords.col=1:2, data.col=4, covar.col=9)
summary(dB)
plot(dB)
points(dB)
binc = variog(dB, trend=~distancia, uvec = seq(0,220000,l=20))
plot(binc,main="com covariável")
bins = variog(dB, tolerance = to, uvec = seq(0,220000,l=20))
plot(bins,main="sem covariável")
# VARIOGRAMA CLOUD, BIN E BOX-PLOTS
par(mfcol = c(3,2), mar=c(3,3,2,2),mgp=c(1.5,.7,0))
cloud1 <- variog(dB, option = "cloud", uvec = seq(0,220000,l=20),trend=~distancia)#, tolerance=
to)
plot(cloud1, main = "Cloud covariável")
vario1 <- variog(dB, uvec = seq(0,220000,l=20), trend=~distancia, tolerance = to)
plot(vario1,main="Bin covariável")
bin1 <- variog(dB, uvec = seq(0,220000,l=20), bin.cloud = T, trend=~distancia, tolerance = to)
#BOX-PLOTS
plot(bin1, bin.cloud = T)
cloud2 <- variog(dB, uvec = seq(0,220000,l=20), option = "cloud")#, tolerance= to)
plot(cloud2, main = "Cloud")
vario2 <- variog(dB, uvec = seq(0,220000,l=20))#, tolerance = to)
plot(vario2,main="Bin")
bin2 <- variog(dB,uvec = seq(0,220000,l=20), bin.cloud = T)#, tolerance= to)
plot(bin2, bin.cloud = T)
# VERIFICANDO AS DIFERENTES DIREÇÕES (ANISOTROPIA)
par(mfrow=c(1,2))
to=pi/18
md=220000
a1=0                # 0°
a2=pi/4             # 45°
a3=pi/2             # 90°
a4=pi/1.333333      # 135°
plot(variog4(dB, uvec = seq(0,md,l=30),direction = c(a1,a2,a3,a4), tolerance= to))
plot(variog4(dB, trend =~distancia, uvec = seq(0,md,l=30),direction = c(a1,a2,a3,a4), tolerance=
to))
# AJUSTE DO MODELO SEM COVARIÁVEL
bins = variog(dB, uvec = seq(0,220000,l=20))#,option = "cloud")       1.79, 166486.49
dB.ef <- eyefit(bins)
ols <- variofit(binS, cov.model = "gauss", fix.nugget = T, nugget=__,max.dist=220000,
weights="equal")
wls <- variofit(binS, cov.model = "gauss", fix.nugget = T, nugget=0.06, max.dist=220000)
```

```
ml <- likfit(dB, cov.model="gauss", fix.nugget = T, nugget=0.06, ini=c(0.12,
124864.86),lik.method = "ML")
reml <- likfit(dB, cov.model="gauss", fix.nugget = T, nugget=0.06, ini=c(0.12,
124864.86),lik.method = "REML")
plot(binc,main = expression(paste("Modelo Gaussiano")))
lines(ols, lty = 1, col="red")
lines(wls, lty = 1, col="blue")
lines(ml, lty = 1, col="orange")
lines(reml, lty = 1, col="black")
legend(170000, 0.05, legend=c("OLS","WLS","ML","REML"), lty=rep(1,4),
col=c("red","blue","orange","black"), cex=1)
# AJUSTE DO MODELO COM COVARIÁVEL
ols <- variofit(binc, cov.model = "gauss", fix.nugget = T, nugget=0.06,max.dist=220000,
weights="equal")
wls <- variofit(binc, cov.model = "gauss", fix.nugget = T, nugget=0.06, max.dist=220000)
binc = variog(dB, trend=~distancia, uvec = seq(0,220000,l=20))#,option = "cloud")        1.79,
166486.49
dB.ef <- eyefit(binc)
ml <- likfit(dB, trend=~distancia, cov.model="gauss", fix.nugget = T, nugget=0.06, ini=c(0.12,
124864.86),lik.method = "ML")
reml <- likfit(dB, trend=~distancia, cov.model="gauss", fix.nugget = T, nugget=0.06, ini=c(0.12,
124864.86),lik.method = "REML")
plot(binc,main = expression(paste("Modelo Gaussiano")))
lines(ols, lty = 1, col="red")
lines(wls, lty = 1, col="blue")
lines(ml, lty = 1, col="orange")
lines(reml, lty = 1, col="black")
legend(170000, 0.05, legend=c("OLS","WLS","ML","REML"), lty=rep(1,4),
col=c("red","blue","orange","black"), cex=1)
ols
wls
ml
reml
summary(ols)
summary(wls)
summary(ml)
summary(reml)
#TESTANDO A COVARIÁVEL
ml$beta
ml$beta.var
#IC (95%) para o coeficiente da covariável
ml$beta[2] +  qnorm(c(0.025, 0.975)) * sqrt(ml$beta.var[2,2])
reml$beta
reml$beta.var
#IC (95%) para o coeficiente da covariável
reml$beta[2] +  qnorm(c(0.025, 0.975)) * sqrt(reml$beta.var[2,2])
# VALIDAÇÃO CRUZADA
xv.ols <- xvalid(dB, model=ols)
```

```
xv.wls <- xvalid(dB, model=wls)
xv.ml <- xvalid(dB, model=ml)
xv.reml <- xvalid(dB, model=reml)
summary(xv.ols)
summary(xv.wls)
summary(xv.ml)
summary(xv.reml)
par(mfcol = c(5,2), mar=c(3,3,.5,.5), mgp=c(1.5,.7,0))
plot(xv.ols)
plot(xv.wls)
plot(xv.ml)
plot(xv.reml)
# MALHA A SER PREDITA
loci <- read.table("malha.txt",header=F)
# KRIGAGEM ORDINÁRIA SEM COVARIÁVEL
kc <- krige.conv(dB, locations =
loci[,1:2],krige=krige.control(obj.model=reml,type.krige="OK"))
loci$pred <- kc$predict
loci$vari <- kc$krige.var
# matrix(c(kc$predict,kc$krige.var),ncol=2,dimnames = list(c(),c("predict", "var")))
write.table(loci,file="reml_cov.txt",row.names=F)
# KRIGAGEM ORDINÁRIA COM COVARIÁVEL
kc <- krige.conv(dB, locations =
loci[,1:2],krige=krige.control(obj.model=reml,type.krige="OK",trend.d=~distancia,trend.l=~loci[
,3]))
loci$pred <- kc$predict
loci$vari <- kc$krige.var
# matrix(c(kc$predict,kc$krige.var),ncol=2,dimnames = list(c(),c("predict", "var")))
write.table(loci,file="reml_cov.txt",row.names=F)
```