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

GEOESTATÍSTICA: ESTIMAÇÃO E PREDIÇÃO SUPONDO UM PROCESSO GAUSSIANO SUBJACENTE

Autor: Gustavo da Silva Ferreira
Orientadora: Jandyra M. G. Fachel

Monografia apresentada para a obtenção
do grau de Bacharel em Estatística

Porto Alegre, Abril de 2002

AGRADECIMENTOS

Ao final destes longos anos de estudos, é necessário agradecer às pessoas que me ajudaram de forma significativa a concluir este curso de graduação.

Agradeço à professora Jandyra pela orientação, empenho e auxílio na elaboração deste trabalho e pelos incentivos a seguir em frente.

Agradeço aos meus professores que, ao longo dos anos, transmitiram o conhecimento à respeito da minha profissão.

Agradeço aos meus pais, Glênio e Graziela, pela educação, paciência, amor e dedicação demonstrados, não só no período do curso, mas ao longo de toda a minha existência.

Agradeço aos meus irmãos, Guilherme e Gisele, pelo carinho e por serem uma motivação para eu alcançar este objetivo.

Agradeço aos colegas que conheci ao longo do curso, pois me acolheram e nos tornamos grandes amigos de confraternizações e importantes parceiros de estudo.

Agradeço à todos que, de alguma forma, colaboraram para que este momento chegasse.

E, acima de tudo, gostaria de agradecer ao Senhor Deus por estar comigo ao longo de toda a minha vida, sempre me capacitando a enfrentar as barreiras e dificuldades encontradas ao longo do caminho.

SUMÁRIO

1. INTRODUÇÃO

A inferência estatística busca tirar conclusões a respeito de um certo fenômeno em estudo com base nos resultados observados em amostras ou realizações deste fenômeno. A inferência estatística fornece-nos subsídios para podermos estimar certas características do fenômeno conhecendo a precisão e a confiança dos resultados obtidos.

O conhecimento do comportamento do fenômeno em estudo é de grande ajuda para a escolha dos estimadores e para chegarmos a resultados precisos. Quando conhecemos este comportamento, podemos estudar a variabilidade de um estimador para algum parâmetro envolvido e, consequentemente, obtermos uma quantificação do erro que se comete ao utilizá-lo no processo de estimação.

Através do Cálculo de Probabilidades, conhecemos diversos modelos de distribuições teóricas que dependem de alguns parâmetros que as caracterizam totalmente. Uma vez que a distribuição teórica do fenômeno em estudo é conhecida e realizamos um experimento, podemos estimar seus parâmetros a fim de buscar esta caracterização.

A tentativa de descrevermos o fenômeno em estudo através de um certo modelo probabilístico nem sempre é uma tarefa fácil, sendo muitas vezes necessária a utilização de métodos empíricos. Quando se possui pouco ou nenhum conhecimento sobre o fenômeno, pode ser um pouco audacioso demais fazer afirmações do tipo: *o fenômeno segue uma distribuição Poisson.* Este problema é também conhecido como problema de especificação (Costa Neto, 1977). Técnicas consagradas da inferência estatística, como os métodos de máxima verossimilhança e a abordagem bayesiana, estão fundamentadas no conhecimento da forma da distribuição de probabilidade do fenômeno em questão.

A Geoestatística, cujas fundamentações teóricas estão baseadas na *Teoria das Variáveis Regionalizadas* (Matheron[1], 1963-1971 apud Cressie, 1991), tem se utilizado das técnicas estatísticas de estimação para estudos onde o fenômeno em questão se distribui espacialmente. Este ramo da ciência possui suas raízes nos estudos de mineração e é atualmente aplicada com frequência por geólogos, biólogos, engenheiros, matemáticos e estatísticos (Cressie, 1991).

---

[1] Matheron, G. Principles of geostatistics. *Economic Geology*, 58 (8): 1246-1266, Dec.1963 e Matheron, G. *The theory of regionalized variables and its applications.* Paris, Les Cahiers du Centre de Morphologie Mathematique de Fontainebleu, 1971. 211p.

Resumidamente, a Geoestatística se caracteriza por análises de um processo estocástico contínuo *d-dimensional* $S(x)$, que se distribui dentro de uma região D, onde $\{S(x): x \in D\}$ e $D \in \Re^d$. Nos estudos de Geoestatística, objetiva-se fazer predições dos valores de $S(x)$, para um dado local x não observado, com base nos valores observados nos locais vizinhos. A partir disto, pode-se distinguir dois grandes enfoques para a realização de predições em Geoestatística: métodos baseados em *Interpolações* e métodos baseados em *Modelos de estrutura de covariância espacial*.

Os métodos baseados em interpolações, tais como a Triangulação, Inverso da distância e Kernel, se preocupam apenas com os efeitos de variação de primeira ordem, ou seja, em como varia a média ou valor esperado do fenômeno ao longo da região em estudo. Tais métodos não levam em consideração os efeitos de segunda ordem, ou seja, não exploram a dependência espacial dos desvios em relação à média. Ao invés disso, simplesmente se baseiam na realização de médias simples ou ponderadas, estas em geral ponderadas pela distância, dos locais observados próximos ao local que se deseja realizar a predição (Bailey & Gatrell, 1995). Estes métodos não levam em consideração nenhum modelo probabilístico e, portanto, produzem estimativas sem quantificar o erro envolvido. São procedimentos empíricos que necessitam de amostras bem representativas a fim de possuírem alguma validade científica. Métodos baseados neste enfoque não serão apresentados nesta monografia.

Por outro lado, a abordagem baseada em modelos de estrutura de covariância espacial supõe modelos estocásticos e se baseia na estimação dos parâmetros deste modelo. Num segundo momento, predições são realizadas segundo este modelo, cujos parâmetros foram estimados através de uma amostra. As técnicas de Krigeagem são normalmente utilizadas nesta fase de predição do fenômeno. Considerações referentes a esta discussão sobre abordagens estatísticas baseadas em modelos e baseadas em métodos empíricos podem ser encontradas em Hansen[2] et al.(1983), apud Cressie (1991).

Dentro do enfoque baseado em modelos de estrutura de covariância espacial, a análise estatística costuma se desenvolver baseada no pressuposto da existência de uma dependência, ou correlação espacial, que um determinado ponto da região D possui com os pontos próximos a ele. Esta dependência espacial é geralmente estimada com base

---

[2] Hansen, M. H., W. G. Madow, and B. J. Tepping. 1983. An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association* 78:776–760.

nos valores de $S(x_i)$, onde $i = 1, 2, ..., n$, correspondem aos n locais observados na amostra.

Entretanto, a história mostra que o desenvolvimento das técnicas de Geoestatística ocorreu de duas formas paralelas e independentes, apesar de equivalentes. De um lado a abordagem desenvolvida por Matheron e seus colegas pesquisadores; de outro, os estudos de Matérn, Whittle, Bartlett.e outros pesquisadores, dentro de um contexto referente à Estatística Espacial (Diggle & Ribeiro, 2000). Posteriormente Brian Ripley[3] (1981), apud Diggle & Ribeiro (2000) fez uma conexão explícita entre as duas abordagens. Mais tarde, Cressie (1991) definiu a Geoestatística como um dos três principais ramos da Estatística Espacial.

Atualmente, as técnicas de estimação e predição em Geoestatística ainda costumam ser utilizadas segundo o contexto da mineração, desenvolvido inicialmente por Matheron. Entre outros aspectos, esta abordagem não faz menção quanto à forma da distribuição de $S(x)$, baseando-se somente na caracterização do fenômeno através de um modelo teórico de dependência espacial entre os locais da região D. Esta falta de especificação do modelo $S(x)$ faz com que o uso de técnicas consagradas de inferência estatística, como os métodos de máxima verossimilhança e a abordagem bayesiana, tenham pouco ou nenhum uso. (Diggle & Ribeiro,2000).

Diggle et.al. (1998) utilizaram a expressão *Geoestatística baseada em modelos* para descrever uma abordagem para problemas geoestatísticos baseada na aplicação de métodos estatísticos sob a suposição de modelos estocásticos explícitos. Posteriormente, Diggle & Ribeiro (2000) utilizaram esta abordagem para descrever as técnicas de estimação e predição normalmente utilizadas em Geoestatística, através da suposição de um modelo estatístico predeterminado.

Nesta monografia, apresentaremos os principais tópicos de Geoestatística sob a visão estatística baseada em modelos predefinidos de uma forma simples e clara, objetivando divulgar este enfoque para estudantes e pesquisadores em geral, que buscam iniciar estudos nesta área. Um segundo objetivo será mostrar algumas desvantagens, raramente declaradas, que ocorrem quando se utilizam os métodos tradicionais de estimação da estrutura de covariância espacial, geralmente baseados em *Ajuste de Curvas*, que não fazem menções quanto à forma da distribuição de $S(x)$. Para isto, será necessário assumir um modelo probabilístico para o processo estocástico $S(x)$.

[3] Ripley, B.D. (1981). *Spatial Statistics*. New York: Wiley.

O modelo que será utilizado supõe que a distribuição de $S(x)$ é Gaussiana Multivariada, com a vantagem de ser de fácil tratamento, além de não se constituir em uma suposição muito forte para uma boa parte de estudos em Geoestatística. Outro importante motivo para a utilização da distribuição Gaussiana para $S(x)$ está no fato de que as principais técnicas em Geoestatística atualmente utilizadas estão, implicitamente, assumindo esta suposição.

Esta monografia terá como base a publicação *Model Based Geostatistics* (Diggle & Ribeiro, 2000). Esta obra contém as principais idéias e conceitos utilizados no desenvolvimento desta monografia. Para a análise e comparação entre os procedimentos apresentados, foram gerados dados através de simulações e também utilizaram-se dados fornecidos pelo projeto MAPEM. Apesar disto, é importante salientar que o objetivo maior nesta monografia não é realizar um estudo de caso nas variáveis utilizadas, mas a análise e comparação das técnicas geoestatísticas quando aplicadas a diferentes tipos de dados. Maiores esclarecimentos sobre o projeto MAPEM são apresentados em anexo.

A monografia está dividida de forma a contemplar os principais procedimentos realizados para a estimação e predição em Geoestatística.

No Capítulo 2 apresentaremos uma visão geral da Geoestatística, partindo de suas origens, com a fundamentação da Teoria das Variáveis Regionalizadas, passando pela descrição do modelo geoestatístico até apresentarmos maneiras de realizar predições através do erro quadrático médio.

O Capítulo 3 é reservado para o estudo da estrutura de covariância espacial de modelos geoestatísticos. Apresentaremos as definições e relações de variograma, covariograma e correlograma, bem como as principais funções de correlação espacial.

No Capítulo 4 apresentaremos maneiras de realizar a escolha de um modelo adequado para a estrutura de covariância espacial, bem como maneiras de estimar os parâmetros envolvidos no modelo escolhido. Apresentaremos também algumas técnicas exploratórias a fim de verificar se as suposições envolvidas no modelo geoestatístico estão satisfeitas. Por fim, realizaremos a estimação da estrutura de covariância espacial para processos simulados e para os dados do projeto MAPEM. Este processo será feito de forma comparativa entre os métodos de estimação baseados no critério da máxima verossimilhança e baseados no critério de mínimos quadrados.

No Capítulo 5 apresentaremos a ligação entre o processo de estimação da estrutura de covariância espacial e o processo de predição espacial. Através das

simulações e dos dados fornecidos pelo projeto MAPEM, realizaremos a predição supondo diferentes modelos. Os modelos utilizados para a predição serão os mesmos modelos estimados no Capítulo 4 pelos critérios da máxima verossimilhança e mínimos quadrados.

No Capítulo 6 apresentaremos introdutoriamente o pacote computacional GeoR, o qual foi utilizado na monografia a fim de realizar os procedimentos de análise, comparações de técnicas e geração de dados geoestatísticos. Neste capítulo, serão apresentados os principais comandos do pacote GeoR necessários para a realização de uma análise Geoestatística.

A contribuição desta monografia está no fato de que grande parte dos livros de Geoestatística não são direcionados aos estatísticos. Normalmente, seu público alvo são os pesquisadores dos principais fenômenos onde se aplicam a Geoestatística, ou seja, geólogos, engenheiros, biólogos, etc. Outra dificuldade, é o fato de existirem poucos textos em língua portuguesa que tratam deste assunto.

Em vista disso, creio que as considerações desta monografia devam servir como auxílio para o início dos estudos em Geoestatística por parte de estudantes de estatística e pesquisadores em geral.

2. A TEORIA DAS VARIÁVEIS REGIONALIZADAS E A GEOESTATÍSTICA

2.1. A TEORIA DAS VARIÁVEIS REGIONALIZADAS

A variabilidade espacial, principalmente no que diz respeito às características do solo, vem sendo uma das principais preocupações dos pesquisadores desde o início do século. A necessidade de se criar métodos para a análise de dados que se distribuem no espaço ou no tempo surgiu do fato que estes fenômenos exibem comportamentos demasiadamente complexos para serem analisados pelos métodos estatísticos usuais.

No início da década de 50, Daniel G. Krige[4] (1951), apud Cressie (1991), trabalhando com dados de concentração de ouro, concluiu que somente a informação dada pela variância seria insuficiente para explicar o fenômeno em estudo. Para tal, seria necessário levar em consideração a distância entre as observações. Krige foi o pioneiro em introduzir o uso de médias móveis para evitar superestimação sistemática de reservas em mineração (Cressie, 1991). A partir daí surgiam os fundamentos da Geoestatística, que leva em consideração a localização geográfica e a dependência espacial entre locais amostrados no processo de predição espacial.

Fundamentada por Matheron, com base nas observações de Krige, a *Teoria das Variáveis Regionalizadas* pode ser entendida como o estudo de uma variável distribuída no espaço ou tempo cujos valores são considerados como realizações de uma função aleatória ou processo estocástico.

Sob uma visão matemática, podemos definir uma variável regionalizada da seguinte forma: Seja $x \in \Re^d$ um local no espaço euclidiano d-dimensional e seja $S(x)$ uma variável que assume um valor aleatório para um dado local x. Com x variando sobre todo um espaço $D \in \Re^d$, teremos gerado um processo estocástico multivariado

$$\{S(x) : x \in D\} \tag{2.1}$$

As primeiras aplicações desta teoria deram-se no contexto da mineração, onde buscava-se predizer os valores de certas características minerais do solo através de alguns pontos amostrados em uma área predeterminada. Com o passar dos anos, o

[4] Krige, D.G. (1951) *A statistical approch to some basic mine evaluation problems on the Witwatersrand*. Johanesburg Chemistry Metallurgy Mining Society South African, 52 (6): 119-139, 1951.

fundamento desta técnica foi aplicado em diversas áreas do conhecimento humano, tais como demografia, agronomia, epidemiologia, ecologia, sociologia, etc. Isto originou um ramo aplicado da estatística chamado de *Estatística Espacial*.

A Estatística Espacial considera os valores amostrais como sendo realizações de funções aleatórias com distribuição no espaço e, nesse caso, o valor de um ponto é função da sua posição na região de estudo. Outro fator que também é levado em consideração na estatística espacial é a posição relativa dos pontos amostrados. Assim, a similaridade entre valores amostrais é quantificada em função da distância entre amostras, representando tal relação o fundamento desse campo especial da estatística aplicada.

Podemos distinguir três grandes subdivisões de problemas em Estatística Espacial (Cressie, 1991): Geoestatística, Dados de Área e Padrões de Pontos.

Geoestatística

A Geoestatística, cuja estrutura teórica foi apresentada na teoria das variáveis regionalizadas, é aplicada nos casos em que supomos um processo estocástico que assume valores sobre *todos* os locais da região d-dimensional $D \in \Re^d$. A Geoestatística se fundamenta no reconhecimento da variabilidade espacial do processo, tanto em grande como em pequena escala. Assim, os modelos baseiam-se na análise da tendência e/ou da correlação espacial. Um exemplo de problema em Geoestatística é considerar uma amostra de medições do nível de poluição em 30 locais no centro de Porto Alegre onde se deseja estimar como estão os níveis de poluição para *todo o centro de Porto Alegre*. A principal característica que diferencia a Geoestatística dos demais tipos de problemas espaciais está no fato de que *x varia continuamente sobre toda a região* $D \in \Re^d$.

Dados de Área

A análise de Dados de Área ou *Lattice Data* (Cressie, 1991) é aplicada nos casos em que cada local s é considerado uma área fixa contendo inúmeros pontos pertencentes à $\Re^d$. Normalmente, cada s é dividido segundo limites geográficos e políticos, tais como municípios, ou segundo alguma divisão que vise algum

planejamento de experimento. Aqui se busca encontrar padrões nos valores de $S(x)$ em relação ao espaço. Diferentemente da Geoestatística, a análise de Dados de Área permite dados que podem ser exaustivos do fenômeno, ou seja, os dados constituem a população já que não há amostragem. Como exemplo, podemos considerar um estudo onde buscamos relações do PIB de cada um dos estados brasileiros com a localização geográfica. Neste caso, a região D de estudo seria o próprio espaço $\Re^d$, ou seja, o território brasileiro. Note que à medida que possuímos os dados do PIB de todos os estados, não desejamos realizar predições, mas sim, buscar relações ou padrões dos valores do PIB no espaço.

Padrões de Pontos

A análise de Padrões de Pontos surge quando se deseja estudar o comportamento no espaço da ocorrência de um certo evento ou fenômeno expresso através de ocorrências pontuais. Definimos um padrão pontual como um conjunto de dados consistindo de uma série de localizações pontuais que indicam a ocorrência de eventos de interesse dentro da área de estudo. O termo *evento* é utilizado de forma geral para referir-se a qualquer tipo de fenômeno localizado no espaço que possa estar associado a uma representação pontual. O objetivo de análises deste tipo de problemas é descobrir se o fenômeno ocorre de maneira *aleatória, regular, agrupada ou temporal em relação ao espaço*. O exemplo típico deste tipo de problema é analisar incidência de uma doença numa determinada região. Através de análises de padrões de pontos, pode-se verificar se a doença está distribuída aleatoriamente no espaço-tempo, se está concentrada em algum local específico, ou ainda, se estamos diante de uma epidemia que está se alastrando ao longo da região.

Nesta monografia, não serão abordados os problemas de análise de Dados de Área e Padrão de Pontos. Para um estudo destes tipos de problemas espaciais, recomenda-se a leitura de *Statistical for Spatial Data* (Cressie, 1991) e *Interactive Spatial Data Analysis* (Bailey & Gatrell, 1995).

2.2. GEOESTATÍSTICA

A Geoestatística é um tópico da Estatística Espacial que possui ampla aplicação em inúmeras áreas do conhecimento. Seu foco central é a análise de dados originários de processos estocásticos distribuídos de forma contínua no espaço, ou seja, de variáveis regionalizadas (ver Figura 2.1).

A idéia básica da Geoestatística — idéia que pode ser considerada como o alicerce de praticamente todas as técnicas deste ramo da Estatística Espacial — é a de que observações vizinhas no espaço tendem a possuir valores de atributo próximos e, à medida que a distância entre os pontos aumenta, esta similaridade diminui gradualmente. Como a variável regionalizada se distribui de forma contínua no espaço, apenas alguns pontos, obtidos através de amostragem, têm o seu valor conhecido. O tamanho, o arranjo espacial e os valores de atributo dessas amostras constituem o suporte da inferência em Geoestatística.

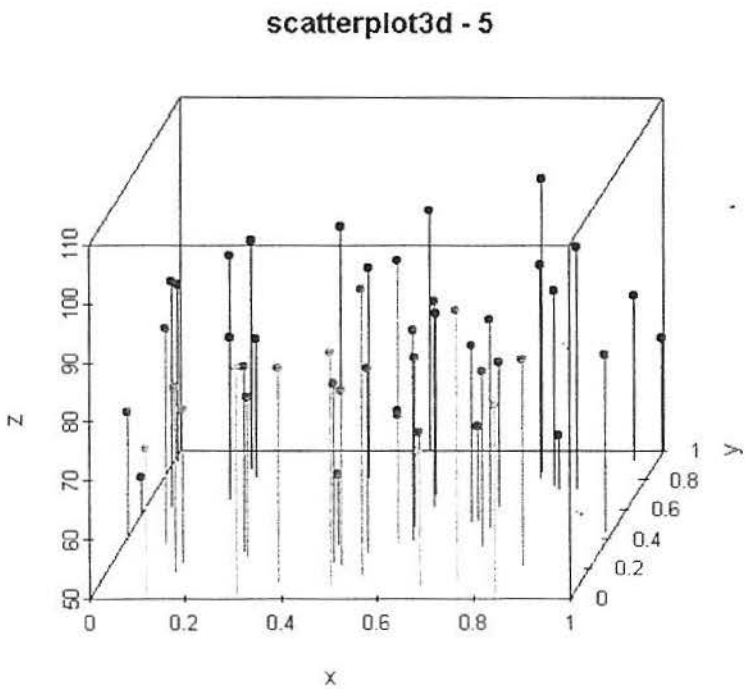

Figura 2.1 - Apresentação gráfica de uma variável regionalizada.

Em Geoestatística, o estudo do modelo de correlação expresso por um conjunto de dados é denominado Análise Estrutural. As inferências para locais não amostrados são realizadas por procedimentos como a krigeagem. De forma resumida, os passos em

um estudo empregando técnicas geoestatísticas incluem: *Análise exploratória dos dados, Análise da estrutura de covariância espacial* e *Predição espacial.*

O emprego destas técnicas visa resolver as seguintes questões:

i) O entendimento, através de funções matemáticas, das leis naturais que governam fenômenos distribuídos no espaço;

ii) Predição de valores das variáveis regionalizadas ou estimação de parâmetros associados às suas características espaciais;

iii) A avaliação dos erros de estimação, a fim de estabelecer o grau de confiabilidade em previsões assegurando que um erro máximo de estimação não será excedido.

Diversos métodos têm sido propostos a fim de inferir valores para locais não amostrados. Estes métodos podem ser classificados em dois grandes grupos:

i) *Métodos baseados nos efeitos de 1ª ordem*: Analisam a *tendência*, ou seja, visam avaliar como varia a média da variável regionalizada ao longo da região em estudo. Estes métodos normalmente baseiam-se na obtenção de médias simples ou ponderadas (geralmente ponderadas pela distância) dos locais observados próximos ao local que se deseja realizar a predição. Dentre estes, podemos destacar a Interpolação Linear por Triangulação, Ponderação pelo Inverso da Distância e Kernel;

ii) *Métodos baseados nos efeitos de 2ª ordem*: São métodos que, adicionalmente, exploram a estrutura de covariância espacial do fenômeno. Estes métodos possuem a vantagem de produzirem estimativas quantificando os erros envolvidos na estimação. Isto ocorre devido a especificação de modelos probabilísticos subjacentes aos dados. Os métodos mais conhecidos deste tipo de análise são os métodos de Krigeagem.

Os métodos que utilizam a abordagem baseada no item ii) são conhecidos como *métodos baseados em modelos* — devido à idéia de utilização de métodos estatísticos formais sob a suposição modelos estocásticos explícitos. Nos métodos de Geoestatística baseados em modelos, normalmente utiliza-se uma ferramenta fundamental a fim de auxiliar no processo de captação da dependência espacial da variável regionalizada: *o variograma*. O variograma tem como principal característica o fato de medir o grau médio de dissimilaridade entre locais x que se encontram igualmente distantes. Nos próximos capítulos trataremos com mais detalhes o variograma e o método de predição conhecido como Krigeagem Simples.

2.2.1. OBJETIVOS DA ANÁLISE GEOESTATÍSTICA

Segundo Diggle & Ribeiro (2000), os objetivos da análise Geoestatística podem ser de dois tipos: *estimação* e *predição*. A *estimação* refere-se a inferência de parâmetros que definem o modelo estocástico gerador dos dados. Estes parâmetros podem ser de interesse científico direto, como os que definem uma regressão que relaciona a variável resposta e alguma variável explicativa, ou de interesse indireto, como aqueles que definem a estrutura de covariância do modelo *S(x)*. A *predição* refere-se à inferência de realizações de *S(x)* em locais não observados. Geralmente a predição objetiva estimar os valores de *S(x)* para todos os locais da sub-região D ($D \in \Re^d$), fornecendo como resultado um mapa de superfície que cobre todos os locais x, amostrados ou não, da região estudada.

2.2.2. O MODELO GEOESTATÍSTICO

Antes de definirmos um modelo de função de distribuição para *S(x)*, precisamos de algumas considerações adicionais sobre o modelo geoestatístico.

O formato básico de dados univariados geoestatísticos é da forma:

$$(x_i, y_i) : i = 1,\ldots,n \qquad (2.2)$$

onde x_i identifica o local no espaço e y_i é uma medida escalar obtida no local x_i. Normalmente x_i é um vetor de coordenadas no espaço bidimensional. Entretanto, exemplos de casos unidimensionais e tridimensionais também ocorrem na prática. A princípio, a medida y_i pode ocorrer em qualquer local da região de estudo D. Os locais x_i podem ter sido amostrados de forma determinística — quando os x_i formam uma espécie de malha sobre a região D — ou de forma aleatória, *independente* do processo que gerou as medidas y_i (ver Figura 2.2).

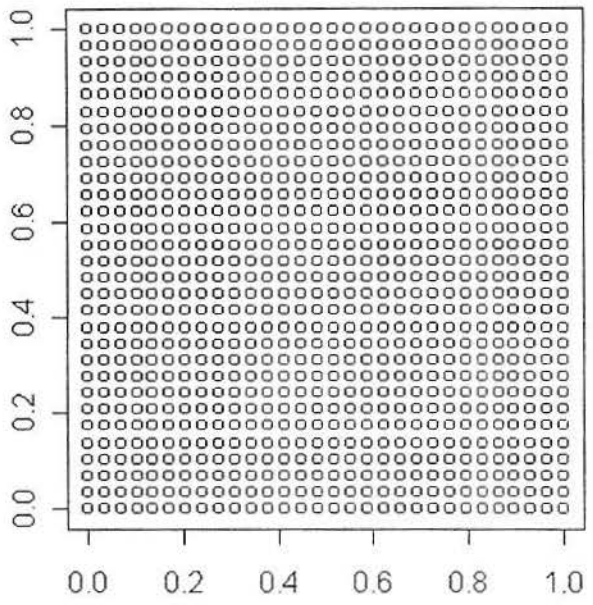

Figura 2.2 - Exemplo de região $D \in \Re^2$ por onde x percorre.

Segundo Diggle & Ribeiro (2000), o modelo geoestatístico é definido através de um processo estocástico $\{Y(x): x \in D\}$, considerado como uma parcial realização do processo $\{Y(x): x \in \Re^2\}$. Os Y_i podem ser considerados como uma *versão ruído* do processo estocástico subjacente $S(x_i)$ que, por sua vez, é o valor do processo $\{S(x): x \in \Re^2\}$ no local x_i. Este modelo básico pode ser ampliado para um modelo mais geral composto das seguintes características:

i) Um processo estocástico $S(x)$;

ii) Um modelo estatístico para as medidas $\mathbf{Y} = (Y_1,..., Y_n)$ condicional a $\{S(x): x \in \Re^2\}$.

2.2.3. PROCESSOS ESTACIONÁRIOS

Na maioria dos problemas práticos em que as técnicas de Geoestatística se aplicam, confrontamo-nos com uma limitação nos dados que exige a introdução de suposições adicionais no modelo geoestatístico.

A principal limitação com a qual nos deparamos é o fato de possuirmos apenas uma única realização do processo estocástico, ou seja, nossos pontos amostrados formam "*uma amostra constituída de n amostras de tamanho 1 (um)*".

Por este motivo, não podemos obter a função de distribuição de $S(x)$ e, consequentemente, as distribuições de $S(x_i)$, que representam as distribuições marginais de $S(x)$ nos pontos x_i.

Uma forma que é amplamente utilizada para resolver este problema é utilizar os dados recolhidos em toda a região D para estimar a função de distribuição de $S(x_i)$, ou seja, assumimos que o comportamento da função de distribuição *global* é idêntico ao comportamento da função de distribuição *local*.

A utilização desta abordagem para a resolução deste impasse na Geoestatística fazendo com que possamos realizar as inferências a respeito da distribuição do nosso processo estocástico exige que utilizemos uma restrição estacionária, semelhante em concepção à ergodicidade nas séries de tempo dependentes.

Antes de prosseguir, é importante definir os principais níveis de estacionariedade em processos estocásticos no contexto da Geoestatística. Um processo estocástico do tipo

$$\{S(x): x \in D\} \qquad , D \in \Re^d \tag{2.3}$$

possui distribuição *finita-dimensional* normalmente definida por

$$F_{x_1,\ldots,x_m}(s_1,\ldots,s_m) = P\{S(x_1) \le s_1,\ldots,S(x_m) \le s_m\}, \qquad m \ge 1$$

e precisa satisfazer certas condições de simetria e consistência (Cressie, 1991). De uma forma mais geral, a $F_{x_1,\ldots,x_m}(s_1,\ldots,s_m)$ precisa satisfazer as condições de simetria e consistência de Kolmogorov, ou seja, a F precisa ser invariante quando os s_i e os x_i são sujeitos à permutação e

$$F_{x_1,\ldots,x_{k+1}}(s_1,\ldots,s_k,\infty) = F_{x_1,\ldots,x_k}(s_1,\ldots,s_k).$$

Supondo que $\mu(x) = E[S(x)]$ existe para todos os locais $x \in D$ — média na qual chamaremos de *tendência* ou *drift* — e supondo a existência de $Var[S(x)]$ para todos $x \in D$, podemos definir três tipos de estacionariedade: *de segunda ordem*, *estrita* e *intrínseca*.

Estacionariedade de Segunda-Ordem

Supondo que a média do processo é constante para todos os locais x da região D, ou seja,

$$E[S(x)] = \mu, \quad \text{para todo } x \in D,$$

e que $F_x(s) = P\{S(x) \leq s\}$ não depende de s, podemos estimar preditores lineares ótimos adicionando a seguinte suposição:

$$Cov[S(x_1), S(x_2)] = C(x_1 - x_2), \quad \text{para todo } x_1, x_2 \in D, \tag{2.4}$$

onde a função $C(\cdot)$ é chamada *covariograma* ou *função covariância estacionária*. Em outras palavras, a covariância do processo $S(x)$ entre dois locais quaisquer depende *apenas* da distância (em módulo e direção) existente entre eles. Este tipo de estacionariedade é também chamado de *estacionariedade fraca* (Cressie, 1991).

Estacionariedade Estrita

Existe um outro tipo de estacionariedade chamada de *estacionariedade forte ou estrita*, que é definida pela relação:

$$F_{x_1,\ldots,x_m}(s_1,\ldots,s_m) = P\{S(x_1) \leq s_1,\ldots,S(x_m) \leq s_m\} = P\{S(x_1+h) \leq s_1,\ldots,S(x_m+h) \leq s_m\}$$

onde $m \geq 1$ e h é um vetor módulo e direção ($h \in \Re^d$). Em outras palavras, dizemos que um processo é fortemente estacionário quando sua lei de distribuição de probabilidade é invariante com a translação. Este tipo de estacionariedade também implica estacionariedade de segunda ordem quando os dois primeiros momentos da $F_x(s)$ são finitos.

Estacionariedade Intrínseca

Podemos também impor condições menos restritivas quanto à estacionariedade dos fenômenos, permitindo que haja maior tolerância para o uso de certas técnicas no tratamento dos mesmos. Desta forma, para um processo estocástico definido como (2.3) temos que a *estacionariedade intrínseca* é definida por:

$$E[S(x+h) - S(x)] = 0,$$

$$Var[S(x+h) - S(x)] = 2\gamma(h)$$

onde h é um vetor módulo e direção ($h \in \Re^d$) e a quantidade $2\gamma(h)$ é conhecida como *variograma*. O variograma é de vital importância e se constitui a principal ferramenta para todo o processo de estimação e predição na análise de fenômenos em Geoestatística. No próximo capítulo definiremos o variograma mais formalmente, apresentando suas vantagens e aplicações.

A estacionariedade intrínseca basicamente impõe que apenas os acréscimos espaciais (h) sejam estacionários.

2.2.4. ISOTROPIA

Quando um processo estocástico estacionário de segunda ordem definido como (2.3), cujo covariograma $C(x_1 - x_2)$ depende apenas de $\|x_1 - x_2\|$, ou seja, da distância euclidiana entre os dois locais, o processo é dito estacionário e *isotrópico*. Voltaremos a tratar de isotropia no próximo capítulo quando tratarmos de variogramas.

2.2.5. ERGODICIDADE

As restrições de estacionariedade têm um papel fundamental em Geoestatística. Isto acontece porque, na maior parte dos fenômenos estudados, dispomos de apenas uma realização do processo gerador dos dados e, consequentemente, não podemos inferir sobre a forma da distribuição de $S(x)$.

Por outro lado, ao assumirmos estacionariedade no nosso processo, podemos estimar seus primeiros momentos através da propriedade ergódica. Ergodicidade não será definida formalmente aqui, mas basicamente a propriedade ergódica ocorre quando a média e a covariância, estimadas a partir de um conjunto restrito de valores, fornecem estimativas não tendenciosas para o conjunto total de valores, requerendo para isto que observações suficientemente distantes umas das outras sejam praticamente não-correlacionadas (ver Harvey, 1981). Uma série temporal estacionária é sempre ergódica com respeito à sua média e com respeito à sua função de autocovariância.

A suposição de *ergodicidade*, no contexto de séries temporais, é de grande utilidade quando dispomos de uma série infinita de dados no tempo $Z(1)$, $Z(2)$, ... , de onde possuímos n observações $z(1)$, $z(2)$,..., $z(n)$. A partir disto, supondo ergodicidade, podemos estimar de maneira consistente a lei de probabilidade de vários subconjuntos $Z(t_1)$, $Z(t_2)$, ..., $Z(t_p)$ (Cressie, 1991).

No contexto de séries temporais é recomendado que, ao se fazer a suposição de ergodicidade, utilizemos o fato de que a média amostral $\overline{Z}$ e a função de autocovariância $\hat{C}(h)$ *convergem em* L_2 para μ e $C(h)$, respectivamente (uma sequência de variáveis aleatórias $\{X_n\}$ converge para X em L_2 se $E(X_n - X)^2 \to 0$

quando $n \to \infty$). Esta mesma recomendação é feita aos geoestatísticos. O único problema que surge ao tentarmos fazer uso desta propriedade é o fato de não estarmos certos quanto à existência da ergodicidade no nosso processo (Cressie, 1991).

Tanto no contexto de séries temporais quanto no contexto de Geoestatística, o problema de buscar informações à respeito da ergodicidade do processo pode ser minimizado se estivermos frente à um processo estocástico Gaussiano estacionário, como definiremos a seguir.

2.3. O MODELO GAUSSIANO ESTACIONÁRIO

Antes de definir o modelo geoestatístico Gaussiano estacionário, vamos realizar algumas considerações a respeito das vantagens de supor que $S(x)$ é Gaussiano multivariado.

Quando lidamos com processos Gaussianos, isto é, quando o processo gerador dos dados tem distribuição conjunta Gaussiana Multivariada, a estacionariedade de segunda ordem e a estacionariedade forte coincidem. Isto acontece porque um processo Gaussiano é caracterizado por sua média e sua função de covariância. A partir disto, uma condição suficiente para assumirmos ergodicidade (Adler[5], 1981 apud Cressie, 1991) é que

$$C(h) \to 0, \text{ quando } \|h\| \to \infty .$$

Além deste fato, quando trabalhamos com processos Gaussianos, todo o processo de predição, estimação e teoria de distribuições é trabalhado analiticamente de maneira "simples". Outra vantagem é o fato de que efeitos de pequena escala que estão presentes nos fenômenos em estudo, acabam por ter distribuição aproximadamente normal devido ao teorema central do limite (Cressie, 1991).

Para definirmos o modelo geoestatístico Gaussiano, primeiramente precisamos realizar algumas considerações sobre o processo $S(x)$.

Um processo estocástico $S(x)$ é Gaussiano se a distribuição conjunta de $S(x_1)$, $S(x_2)$, ..., $S(x_n)$ é Gaussiana multivariada para qualquer inteiro n e qualquer conjunto de locais x_i. Assumindo *estacionariedade de segunda ordem*, a média de $S(x)$ é igual para todo $x \in \Re^2$, e a função de correlação entre $S(x)$ e $S(x')$ depende apenas de

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

$x - x'$. Caso a função de correlação dependa apenas de $\|x - x'\|$, ou seja, a distância euclidiana entre os locais x e x', então o processo é dito ser também *isotrópico* (Diggle & Ribeiro, 2000).

Considerando um conjunto de dados (x_i, y_i), $i = 1, ..., n$, o *Modelo Geoestatístico Gaussiano Estacionário* é definido pelas seguintes suposições (Diggle & Ribeiro, 2000):

i) $\{S(x): x \in \Re^2\}$ é um processo Gaussiano com média μ, variância σ^2 e função de correlação $\rho(h) = Corr\{S(x), S(x')\}$, onde $h = \|x - x'\|$;

ii) Os y_i são realizações mutuamente independentes e identicamente distribuídas de Y_i condicionalmente à $\{S(x): x \in \Re^2\}$, com média condicional $E[Y_i / S(\cdot)] = S(x_i)$ e $Var[Y_i / S(\cdot)] = \tau^2$.

O modelo Gaussiano estacionário pode ainda ser definido de forma equivalente dada por (Diggle & Ribeiro, 2000):

$$Y_i = S(x_i) + Z_i, \qquad i = 1, \ldots, n \tag{2.5}$$

onde $\{S(x): x \in \Re^2\}$ é definido como em i) acima e os Z_i são variáveis aleatórias independentes e identicamente distribuídas N(0, τ^2).

Considerações adicionais devem ser feitas para tornar o modelo válido. Uma delas é a de que a função de correlação $\rho(h)$ precisa ser não-negativa. Isto se faz necessário, pois se a função de correlação assumir valores negativos, poderemos ter valores do variograma $2\gamma(h)$ negativos. Isto pode ser explicado devido a uma relação existente entre o *variograma, o covariograma e o correlograma* que será apresentada no capítulo seguinte.

Existem uma série de famílias paramétricas de funções de correlação que são sugeridas para aplicações em problemas de Geoestatística. Algumas das principais famílias de funções de correlação serão apresentadas no capítulo seguinte, como a Exponencial Potência, a Esférica e a Matérn.

2.4. PREDIÇÃO SUPONDO UM MODELO GAUSSIANO ESTACIONÁRIO

Se quisermos predizer o valor de uma variável aleatória T através de um conjunto de dados y_i, que por sua vez são uma realização de um vetor aleatório $\mathbf{Y}$, então o preditor de T será qualquer função de $\mathbf{Y}$. Chamando o preditor de $\hat{T} = t(\mathbf{Y})$, procuramos encontrar um critério para definir qual a melhor função $t(\cdot)$ a ser escolhida. Um dos critérios mais utilizados em Estatística é o do *erro quadrático médio mínimo*, isto é, escolhemos o valor de $t(\cdot)$ que torne

$$EQM(\hat{T}) = E[(\hat{T} - T)^2]$$

o mínimo possível. Vale lembrar que a esperança dada na expressão acima é com relação à distribuição conjunta de T e $\mathbf{Y}$.

A solução que minimiza a expressão é encontrada a partir do seguinte teorema:

Teorema 2.1: Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *um conjunto de n variáveis aleatórias e* $\mathbf{y} = (y_1, \ldots, y_n)$ *seus valores observados. Seja T qualquer outra variável aleatória e considere* $\hat{T} = t(\mathbf{Y})$ *qualquer função de* $\mathbf{Y}$. *Então o* $EQM(\hat{T})$ *assume seu valor mínimo quando* $\hat{T} = E(T / \mathbf{Y})$.

Prova

Através do cálculo de probabilidades, podemos escrever

$$EQM(\hat{T}) = E[(\hat{T} - T)^2] = E_Y\{E_T[(\hat{T} - T)^2 / \mathbf{Y}]\}. \qquad (2.6)$$

A partir da relação $E(X^2) = Var(X) + [E(X)]^2$, podemos escrever o termo situado dentro da esperança no lado direito da equação (2.6) como

$$E_T[(\hat{T} - T)^2 / \mathbf{Y}] = Var_T[(\hat{T} - T) / \mathbf{Y}] + \{E_T[(\hat{T} - T) / \mathbf{Y}]\}^2.$$

Condicional à $\mathbf{Y}$, qualquer função de $\mathbf{Y}$ na expressão acima é considerada constante. Como $\hat{T}$ é função de $\mathbf{Y}$, podemos eliminá-lo da expressão da variância e ficamos com

$$E_T[(\hat{T} - T)^2 / \mathbf{Y}] = Var_T(T / \mathbf{Y}) + \{E_T(T / \mathbf{Y}) - \hat{T}\}^2$$

Agora aplicando a esperança em relação à $\mathbf{Y}$ nos dois lados da expressão, chegamos à seguinte expressão para o $EQM(\hat{T})$:

$$E_Y\{E_T[(\hat{T} - T)^2 / \mathbf{Y}]\} = E_y[Var_T(T / \mathbf{Y})] + E_y[\{E_T(T / \mathbf{Y}) - \hat{T}\}^2]$$

$$E[(\hat{T} - T)^2] = E_y[Var_T(T / \mathbf{Y})] + E_y[\{E_T(T / \mathbf{Y}) - \hat{T}\}^2]$$

Analisando a expressão acima, vemos que o primeiro termo do lado direito não depende da escolha de $\hat{T}$, enquanto que o segundo termo assume o valor mínimo quando $\hat{T} = E(T / \mathbf{Y})$.

Logo, se usamos $\hat{T} = E(T / \mathbf{Y})$, o $EQM(\hat{T})$ fica reduzido à

$$E[(\hat{T} - T)^2] = E_y[Var_T(T / \mathbf{Y})].$$

A $Var_T(T / \mathbf{Y})$ é chamada de *variância de predição*. O valor da variância de predição, para determinados valores observados y, estima o $EQM(\hat{T})$.

Também é importante notar que em geral $E[(\hat{T} - T)^2] < Var(T)$, ocorrendo a igualdade das duas expressões somente quando T e Y são independentes. Isto ocorre porque $Var(T) = EQM(\hat{T})$ quando $\hat{T} = E(T)$, ignorando a informação obtida através dos dados $\mathbf{Y}$. A diferença entre $Var(T)$ e $Var_T(T / \mathbf{Y})$ pode ser entendida como o ganho na predição de T através da utilização dos dados $\mathbf{Y}$.

Erro Quadrático Médio mínimo de predição em um processo Gaussiano

Vamos agora supor que os dados $\mathbf{Y} = (Y_1, \ldots, Y_n)$ foram gerados pelo modelo Gaussiano estacionário definido como (2.5). Escrevendo $\mathbf{S} = \{S(x_1), \ldots, S(x_n)\}$ para os valores do processo subjacente nos locais $x_1, \ldots, x_n$, onde $\mathbf{S}$ é Gaussiano multivariado com vetor de médias $\mu\mathbf{1}$ (sendo $\mathbf{1}$ um vetor cujos elementos são iguais à 1) e matriz de variâncias $\sigma^2 R$, onde $\mathbf{R}$ é uma matriz n x n com elementos $r_{ij} = \rho(\|x_i - x_j\|)$. De forma similar, $\mathbf{Y}$ é Gaussiano Multivariado com vetor de médias $\mu\mathbf{1}$ e matriz de variâncias $\mathbf{V} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}$, onde $\mathbf{I}$ é a matriz identidade (Diggle & Ribeiro, 2000).

Suponha também que nosso objetivo é predizer o valor de $S(x)$ em um local arbitrário, ou seja, temos que $T = S(x)$. Como $(T, \mathbf{Y})$ é também Gaussiano multivariado, nós podemos obter o $EQM(\hat{T})$ usando um resultado padrão da distribuição Gaussiana multivariada (Chatfield & Collins[6], 1980 apud Diggle & Ribeiro, 2000).

[6] Chatfield, C. and Collins, A.J. (1980). *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Teorema 2.2: Considere $\mathbf{X} = (\mathbf{X}_1 , \mathbf{X}_2)$ *tendo distribuição conjunta Gaussiana multivariada com vetor de médias* $\boldsymbol{\mu} = (\mu_1, \mu_2)$ *e matriz covariância*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

ou seja, $\mathbf{X} \sim MVN(\boldsymbol{\mu}, \Sigma)$. *Então a distribuição de* $\mathbf{X}_1 / \mathbf{X}_2$ *também é Gaussiana multivariada, isto é,* $\mathbf{X}_1 / \mathbf{X}_2 \sim MVN(\boldsymbol{\mu}_{1/2}, \Sigma_{1/2})$ *onde*

$$\boldsymbol{\mu}_{1/2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \tag{2.7}$$

$$\Sigma_{1/2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

—

Aplicando este resultado ao nosso contexto, (T, $\mathbf{Y}$) é Gaussiano Multivariado com vetor de médias $\mu\mathbf{1}$ e matriz de variâncias

$$\begin{bmatrix} \sigma^2 & \sigma^2\mathbf{r}' \\ \sigma^2\mathbf{r} & \tau^2\mathbf{I} + \sigma^2\mathbf{R} \end{bmatrix}$$

onde r é um vetor com elementos $r_i = \rho(\|x - x_i\|)$, onde $i = 1, ..., n$.

Substituindo no Teorema 2.2 as variáveis ($\mathbf{X}_1$, $\mathbf{X}_2$) por (T , $\mathbf{Y}$), temos que o $EQM(\hat{T})$ mínimo para $T = S(x)$ é

$$\hat{T} = \mu + \sigma^2\mathbf{r}'(\tau^2\mathbf{I} + \sigma^2\mathbf{R})^{-1}(\mathbf{Y} - \mu\mathbf{1}) \tag{2.8}$$

com variância de predição

$$Var(T / \mathbf{Y}) = \sigma^2 - \sigma^2\mathbf{r}'(\tau^2\mathbf{I} + \sigma^2\mathbf{R})^{-1}\sigma^2\mathbf{r}. \tag{2.9}$$

Para uma melhor visualização e entendimento das expressões (2.8) e (2.9), podemos escrevê-las da seguinte forma:

$$\hat{S}(x) = \mu + \sigma^2\left[\rho(\|x - x_1\|) \quad \cdots \quad \rho(\|x - x_n\|)\right] \times$$

$$\times \left\{ \left(\tau^2 \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} \rho(\|x_1 - x_1\|) & \cdots & \rho(\|x_1 - x_n\|) \\ \vdots & \rho(\|x_i - x_j\|) & \vdots \\ \rho(\|x_n - x_1\|) & \cdots & \rho(\|x_n - x_n\|) \end{bmatrix} \right) \right\}^{-1} \times$$

$$\times \left(\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} - \mu \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)$$

$$Var[\hat{S}(x)/\mathbf{Y}] = \sigma^2 - \sigma^2\left[\rho(\|x-x_1\|) \quad \cdots \quad \rho(\|x-x_n\|)\right]\times$$

$$\times\left\{\left(\tau^2\begin{bmatrix}1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1\end{bmatrix}+\sigma^2\begin{bmatrix}\rho(\|x_1-x_1\|) & \cdots & \rho(\|x_1-x_n\|)\\ \vdots & \rho(\|x_i-x_j\|) & \vdots\\ \rho(\|x_n-x_1\|) & \cdots & \rho(\|x_n-x_n\|)\end{bmatrix}\right)\right\}^{-1}\times$$

$$\times\sigma^2\begin{bmatrix}\rho(\|x-x_1\|)\\ \vdots\\ \rho(\|x-x_n\|)\end{bmatrix}$$

Através destas duas últimas expressões, podemos entender melhor o papel da estrutura de covariância espacial no processo de predição de $S(x)$.

É importante notar que a variância de predição não depende de **Y**, ou seja, o erro quadrático médio obtido é, consequentemente, igual a variância predita, pois neste caso temos

$$EQM(\hat{T}) = E_Y[Var(T/\mathbf{Y})] \xrightarrow{\text{não depende de Y}} = Var(T/\mathbf{Y}).$$

Entretanto, este resultado é uma das muitas propriedades especiais da distribuição Gaussiana multivariada e não um resultado geral (Diggle & Ribeiro, 2000).

Na terminologia Geoestatística convencional, a construção de uma superfície $\hat{S}(x)$ para todos os locais $x \in D$, onde D representa alguma região qualquer pertencente à $\Re^2$, é chamada de *Krigeagem Simples.*

Erro Quadrático Médio mínimo de predição em uma função linear de um processo Gaussiano

Suponha que ao invés de estarmos interessados na predição de um único valor de $S(x)$, desejamos agora predizer uma função linear de $S(x)$ definida por

$$T = \int_D w(x)S(x)dx$$

onde a função $w(x)$ descreve algum tipo de ponderação. Em virtude da esperança ser um operador linear temos que, para quaisquer valores de **Y**,

$$E[T/\mathbf{Y}] = \int_D w(x)E[S(x)/\mathbf{Y}]dx \qquad (2.10)$$

que pode ser escrito como

$$\hat{T} = \int_D w(x)\hat{S}(x)dx.$$

Este resultado nos mostra que quando desejamos predizer valores de funções lineares de $S(x)$, é suficiente predizer os valores de $S(x)$ para uma determinada região D e avaliar as propriedades de interesse *diretamente* da superfície predita $\{\hat{S}(x) : x \in D\}$.

Além disso, em virtude de estarmos tratando com um processo Gaussiano estacionário, $(T, \mathbf{Y})$ é Gaussiano multivariado e a distribuição preditiva de T é Gaussiana univariada com média dada por (2.10) e variância

$$Var(T / \mathbf{Y}) = \int_D \int_D w(x)w(x')Cov\{S(x), S(x')\}dxdx'$$

Cabe salientar que estes resultados não são válidos para funções não-lineares de $S(x)$.

3. ESTRUTURA DE COVARIÂNCIA ESPACIAL

Em Estatística elementar, a definição de covariância está associada com uma medida da extensão da variação conjunta de duas variáveis. O mesmo ocorre no contexto espacial, exceto que a covariância que medimos não é feita entre duas variáveis, mas em relação à mesma variável em dois locais diferentes. Em virtude das idéias de dependência espacial, esperamos que a covariância entre dois locais — separados por um vetor h — seja maior para distâncias curtas do que para grandes distâncias.

Estas idéias são importantes a fim de definirmos a estrutura de covariância espacial em relação ao modelo Gaussiano apresentado no capítulo anterior. Serão apresentados neste capítulo os aspectos principais da estrutura de covariância espacial para modelos utilizados em Geoestatística. Também serão apresentados os conceitos de *variograma*, *covariograma* e *correlograma*, bem como algumas famílias paramétricas de funções de covariância.

3.1. VARIOGRAMA, COVARIOGRAMA E CORRELOGRAMA

Suponha um processo Gaussiano $\{S(x): x \in \Re^2\}$ como definido no capítulo anterior. A especificação completa deste processo ocorre após identificarmos os efeitos de 1ª e 2ª ordem (Diggle & Ribeiro, 2000).

Os efeitos de primeira ordem são aqueles relacionados com a *função média* $\mu(x) = E[S(x)]$, também chamada de *tendência*. Os efeitos de segunda ordem são aqueles relacionados com a *função de covariância* ou *covariograma* $Cov(x, x') = Cov\{S(x), S(x')\}$. Para uma melhor compreensão é importante definir também a variância $\sigma^2(x) = Var\{S(x)\}$.

A análise dos efeitos de primeira ordem não são o foco central do nosso estudo, pois estamos preferivelmente tratando de processos estacionários. Entretanto, alguns procedimentos para tratar dos efeitos de primeira ordem serão mencionados no capítulo seguinte.

Um processo Gaussiano $S(x)$ estacionário de segunda-ordem[7] e isotrópico possui função média constante $\mu(x) = E[S(x)] = \mu$ e função de covariância que depende somente da distância entre os locais x, ou seja, $Cov(x, x') = Cov\{S(x), S(x')\} = C(\|x - x'\|) = C(h)$. A variância de $S(x)$ também é constante neste caso e é útil para que possamos escrever a covariância da seguinte forma

$$C(h) = \sigma^2 \rho(h), \tag{3.1}$$

onde $\rho(\cdot)$ é a *função de correlação* ou *correlograma*.

O variograma, o qual foi ligeiramente introduzido na seção anterior, é definido como

$$\gamma(x - x') = \tfrac{1}{2}[Var\{S(x) - S(x')\}]. \tag{3.2}$$

A quantidade $\gamma(x - x')$ na verdade é o semi-variograma, embora o prefixo "semi" seja convencionalmente omitido (Diggle & Ribeiro, 2000). A partir deste ponto nos referiremos ao variograma como $\gamma(x - x')$.

Devido o fato de estarmos tratando de processos estacionários, podemos escrevê-lo da seguinte forma:

$$\gamma(x - x') = \tfrac{1}{2}[Var\{S(x) - S(x')\}]$$
$$\gamma(x - x') = \tfrac{1}{2}[\sigma^2 + \sigma^2 - 2C(x - x')]$$
$$\gamma(x - x') = \sigma^2 - C(x - x')$$

e, considerando que o processo seja isotrópico, temos

$$\gamma(h) = \sigma^2 - C(h). \tag{3.3}$$

A partir da relação $C(h) = \sigma^2 \rho(h)$, temos uma forma do variograma em função da função de correlação:

$$\gamma(h) = \sigma^2 - \sigma^2 \rho(h)$$
$$\gamma(h) = \sigma^2 [1 - \rho(h)]. \tag{3.4}$$

A partir desta relação existente entre o variograma, o covariograma e o correlograma, podemos notar que os três fornecem informações similares, embora de formas diferentes, sobre a dependência espacial do processo $S(x)$.

O covariograma e o correlograma possuem formas equivalentes, embora em escalas diferentes. O covariograma é uma função decrescente que inicia em σ^2 (para

[7] Por simplicidade, a partir deste ponto chamaremos processos estacionários em referência à processos estacionários de segunda-ordem.

$h=0$) e tende à zero com o aumento da distância entre os locais ($h \to \infty$). Já o correlograma inicia em 1 e decai à zero com o aumento de h (Figura 3.1 e Figura 3.2).

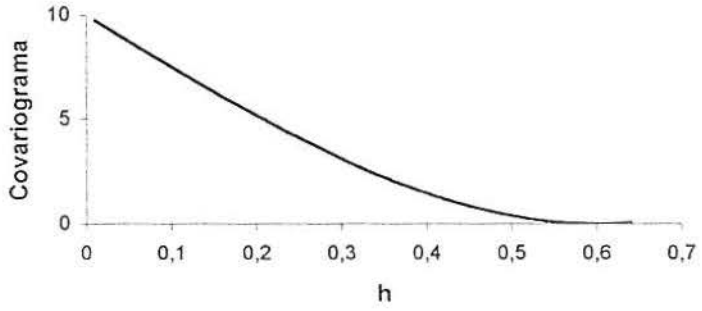

Figura 3.1 - Covariograma ($\sigma^2 = 10$, amplitude = 0,6)

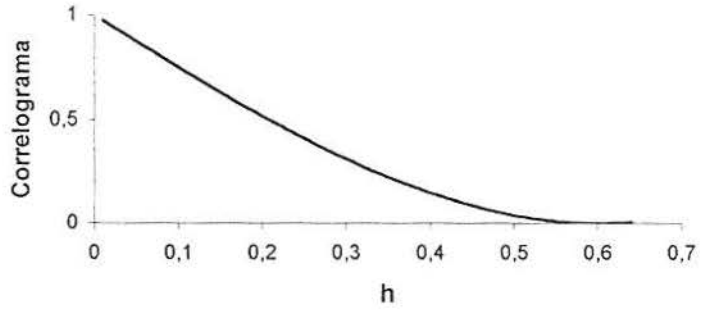

Figura 3.2 - Correlograma ($\sigma^2 = 10$, amplitude = 0,6)

O variograma (Figura 3.3) é semelhante ao covariograma, exceto pelo fato de ser "invertido". O variograma inicia em 0, quando $h = 0$, aumentando até estabilizar em um máximo igual à σ^2, normalmente chamado de patamar (*sill*). O valor de h relativo ao patamar é denominado amplitude (*range*).

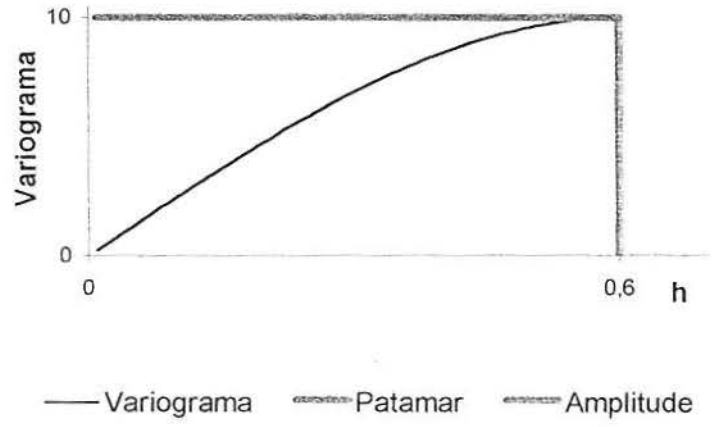

Figura 3.3 - Variograma ($\sigma^2 = 10$, amplitude = 0,6)

Em resumo, o variograma, o covariograma e o correlograma objetivam nos dar uma idéia da relação existente entre os valores de $S(x)$ para dois locais separados por uma distância h. Com a utilização destas informações é possível descobrir a partir de qual distância h, a dependência espacial entre dois pontos torna-se desprezível.

3.2. O EFEITO PEPITA

Considere o modelo Gaussiano definido anteriormente, dado por

$$Y_i = S(x_i) + Z_i \qquad (3.5)$$

onde os Z_i são assumidos como variáveis aleatórias independentes e identicamente distribuídas $N(0, \tau^2)$ e $S(x_i)$ representa o processo Gaussiano estacionário para o local x_i. Neste caso, o termo τ^2 é chamado de *variância pepita*, *efeito pepita* ou, simplesmente, de *pepita*.

O efeito pepita é chamado desta forma devido à sua origem nos estudos de mineração. Nos estudos da variabilidade espacial em campos de mineração de ouro, dois *centros* (locais) de coleta, mesmo que estejam muito próximos, podem gerar resultados consideravelmente diferentes se em um destes locais houver uma pepita de ouro e no outro não. Este fato contraria as idéias da dependência espacial, pois os locais próximos acabam por ter valores de atributo muito diferentes. Ocasiona também implicações no processo de estimação da estrutura de covariância espacial, como poderá ser notado a partir da análise dos textos subsequentes (Journel & Huijbregts, 1978).

Uma das maneiras de interpretar o efeito pepita é considerá-lo como uma *medida de erro*. O fato dos Z_i serem independentes, permite esta interpretação. Para exemplificar, suponha que realizamos duas medições independentes da taxa de radiação Y em uma amostra de n locais x e encontramos variabilidade devida aos dois diferentes instrumentos de medição utilizados. Então a diferença $Y_1 - Y_2$ terá média 0 e variância $2\tau^2$. O valor de τ^2 será a medida da variância do erro.

Outra interpretação dada ao efeito pepita é a de que ele cumpre o papel de modelar os efeitos de variação de pequena escala. Para melhor entendermos esta idéia, vamos supor que o verdadeiro modelo gerador dos dados é

$$Y_i = S(x_i) + T(x_i) + Z_i \,, \tag{3.6}$$

onde $T(x_i)$ é também um processo Gaussiano estacionário independente de Z_i e de $S(x_i)$, cuja variância é σ_T^2 e possui função de correlação $\rho_T(h)$ que decai rapidamente à zero. Suponha que, para $h \geq \alpha$, o valor de $\rho_T(h)$ seja igual à zero. Se o delineamento amostral utilizado para escolher os locais amostrais x_i não selecionar pelo menos 1 (um) par de locais x_i separados por um $h \leq \alpha$, não poderemos distinguir entre o modelo (3.6) e o modelo

$$Y_i = S(x_i) + Z_i^* \tag{3.7}$$

onde os Z_i^* são variáveis i.i.d. $N(0, \tau^2 + \sigma_T^2)$ (Diggle & Ribeiro, 2000).

Ao utilizarmos o modelo (3.7), estaríamos superestimando τ^2, pois não poderíamos diferenciá-lo de σ_T^2, o que nos causaria predições com erros mais elevados. Além disso, devido ao delineamento amostral utilizado, *dificilmente descobriríamos o nosso erro*.

Portanto, vimos que o efeito pepita pode ocorrer devido a erros de medição ou a não especificação correta do modelo, ou seja, à presença de outras variáveis não percebidas pelos pesquisadores. Para minimizarmos este tipo de problema, temos as seguintes sugestões:

i) Sempre que possível, procurar utilizar delineamentos que utilizem repetições de observações em um mesmo local. Assim, teríamos uma boa estimativa do valor de τ^2;

ii) Podemos reservar algumas observações especificamente para estimarmos τ^2. Isto pode ser feito adicionando alguns locais para medição que possuam distâncias pequenas entre si;

iii) Analisar na literatura se existem estudos a respeito do fenômeno desejado e utilizar os resultados destes estudos para auxiliar na estimação de τ^2.

Para melhor explicarmos a relação do efeito pepita com a estrutura de covariância espacial, precisamos apresentar a sua ligação com o variograma.

3.3. O EFEITO PEPITA NO VARIOGRAMA

No caso estacionário, a função de correlação $\rho(\cdot)$ depende apenas de h, ou seja, da distância euclidiana entre os locais x. Este fato implica que $\rho(0)=1$, pois $S(x_i)$ é perfeitamente correlacionado com ele mesmo. Entretanto, quando estudamos um fenômeno em Geoestatística, costumamos trabalhar com uma *versão ruído* de $S(x_i)$, cujo modelo é descrito por

$$Y_i = S(x_i) + Z_i$$

como já mencionado anteriormente. Este modelo incorpora as medidas de erro devidas aos erros de medições, ou seja, ao efeito pepita. Teoricamente, este fato pode fazer com que $\rho(0)<1$. Se nós definirmos um processo $\{Y(x): x \in \Re^2\}$ segundo o modelo descrito em (3.5), onde $x = x_i$, então $Y(x)$ terá variância $\sigma^2 + \tau^2$ e função de covariância $C(h) = \sigma^2 \rho(h)$. Ou seja, a função de correlação é

$$\rho_Y(h) = \frac{\sigma^2 \rho(h)}{\sigma^2 + \tau^2},$$

e, com $h \to 0$, temos que

$$\rho_Y(h) = \frac{\sigma^2 \rho(h)}{\sigma^2 + \tau^2} \xrightarrow{\text{h tendendo à zero}} = \frac{\sigma^2}{\sigma^2 + \tau^2} < 1 .$$

o que explica $\rho(0)<1$.

Por sua vez, o variograma do processo $Y(x)$ é da forma:

$$\begin{aligned}
\gamma_Y(h) &= \tfrac{1}{2}[Var\{Y(x) - Y(x')\}] \\
\gamma_Y(h) &= \tfrac{1}{2}[(\sigma^2 + \tau^2) + (\sigma^2 + \tau^2) - 2\sigma^2 \rho(h)] \\
\gamma_Y(h) &= \tfrac{1}{2}[2\sigma^2 + 2\tau^2 - 2\sigma^2 \rho(h)] \\
\gamma_Y(h) &= \sigma^2 + \tau^2 - \sigma^2 \rho(h) \\
\gamma_Y(h) &= \tau^2 + \sigma^2[1 - \rho(h)] \\
\gamma_Y(h) &= \tau^2 + \gamma_S(h)
\end{aligned}$$

de onde concluímos que o efeito pepita aparece na forma de uma constante adicionada ao variograma, fazendo com que ocorra uma descontinuidade na origem.

Este resultado nos mostra que a variabilidade entre dois locais, separados por uma determinada distância h, é composta de duas partes: uma tratando dos *efeitos de pequena escala* (τ^2) e outra dos *efeitos de grande escala* ($\gamma_S(h)$). Entretanto, esta divisão em duas diferentes causas de variação pode ser estendida para um número maior

de causas de variação, dependendo do fenômeno estudado. Quando temos mais de uma estrutura de variabilidade agindo simultaneamente no nosso processo, elas são chamadas de estruturas *aninhadas* (*Nested Structures*) (Journel & Huijbregts, 1978).

Quando lidados com estruturas aninhadas, o variograma pode ser escrito da seguinte forma:

$$\gamma(h) = \gamma_0(h) + \gamma_1(h) + \gamma_2(h) + \ldots + \gamma_i(h)$$

onde $\gamma_0(h)$ representa o efeito pepita (efeitos de micro-escala), $\gamma_1(h)$ representa os efeitos de pequena escala, $\gamma_0(h)$ representa os efeitos de média escala e $\gamma_i(h)$ representa os efeitos de grande escala. Neste caso, os patamares de cada estrutura seriam diferentes aumentando gradualmente (ver Figura 3.4).

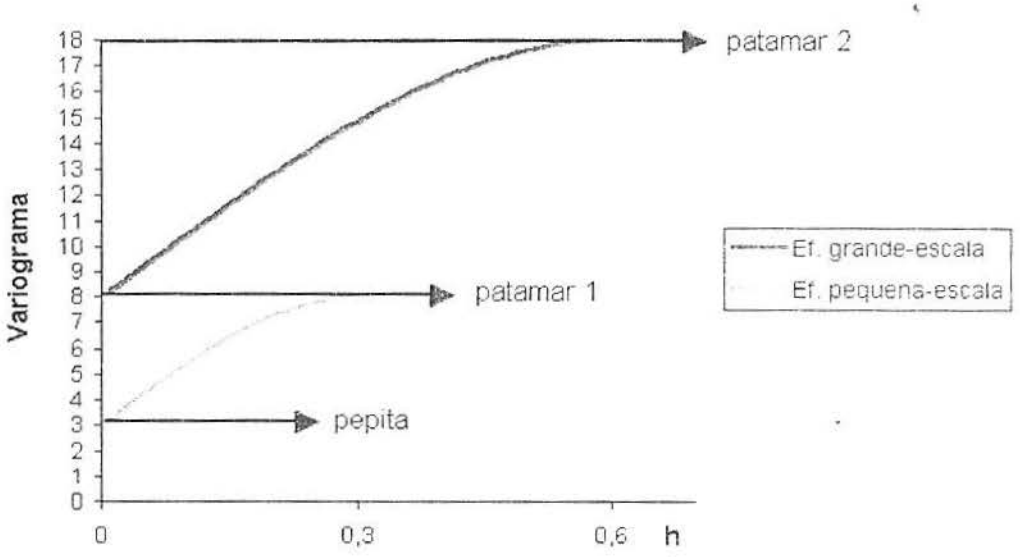

Figura 3.4 - Exemplo de estrutura aninhada

3.4. AUSÊNCIA DE DEPENDÊNCIA ESPACIAL

Quando lidamos com um processo cujo variograma é da forma

$$\gamma(h) = \gamma_0(h) = \tau^2 \qquad , \forall h > 0$$

temos um modelo com ausência de dependência espacial, também chamado de *modelo efeito pepita puro* (Journel & Huijbregts, 1978), cujo variograma está representado na Figura 3.5. Um modelo com efeito pepita puro é semelhante a um ruído branco, onde a única variabilidade existente entre dois pontos separados por uma distância h é causada por efeitos de pequena e micro escala. Quando lidamos com um processo que possui

este tipo de estrutura, a melhor previsão que podemos fazer para $S(x_i)$, seja qual for o local x_i, é *a média* μ do processo.

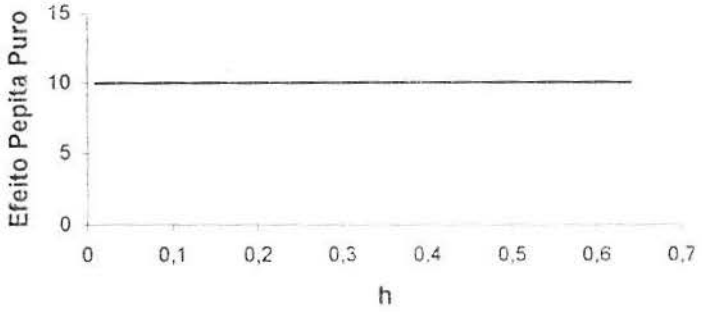

Figura 3.5 - Efeito Pepita Puro ($\tau^2 = 10$)

3.5. DIFERENCIABILIDADE DE PROCESSOS GAUSSIANOS

A diferenciabilidade, no contexto de Geoestatística, nada mais é do que uma descrição do grau de suavização da superfície espacial $S(x)$ (Diggle & Ribeiro, 2000). Para um melhor entendimento desta propriedade, vamos explaná-la para o caso unidimensional x. Um processo $S(x)$ é contínuo em média quadrática *(mean-square continuous)* se, para todo x,

$$E[\{S(x+h) - S(x)\}^2] \to 0, \quad \text{quando } h \to 0.$$

Similarmente, $S(x)$ é diferenciável em média quadrática *(mean-square diferentiable)* se existe um processo $S'(x)$ tal que, para todo x,

$$E\left[\left\{\frac{S(x+h) - S(x)}{h} - S'(x)\right\}^2\right] \to 0, \quad \text{quando } h \to 0.$$

Em outras palavras, um processo $S(x)$ é *diferenciável* em média quadrática se sua derivada quadrática média $S'(x)$ existe. O mesmo ocorre para derivadas de ordem mais altas, ou seja, o processo é *duplamente diferenciável* em média quadrática se existe um processo $S''(x)$ que é a derivada quadrática média de $S'(x)$.

A diferenciabilidade quadrática média de $S(x)$ está diretamente ligada com a diferenciabilidade da função de covariância, como nos diz o seguinte teorema (Diggle & Ribeiro, 2000):

Teorema 3.1: Seja $S(x)$ um processo Gaussiano estacionário com função de correlação $\rho(h): h \in \Re$. Então:

i) *$S(x)$ é contínuo em média quadrática se, e somente se, $\rho(h)$ é contínua nas proximidades de $h = 0$;*

ii) *$S(x)$ é k vezes diferenciável em média quadrática se, e somente se, $\rho(h)$ é no mínimo $2k$ vezes diferenciável em $h = 0$.*

3.6. FAMÍLIAS DE FUNÇÕES DE CORRELAÇÃO

Como vimos anteriormente na Seção 3.3, os variogramas normalmente são da forma

$$\gamma(h) = \tau^2 + \sigma^2[1 - \rho(h)].$$

A função de correlação $\rho(h)$ que aparece na forma do variograma necessita obrigatoriamente ser uma função definida positiva. Se essa condição estiver satisfeita, temos a garantia de que qualquer combinação linear de valores do processo Gaussiano $S(x)$, como definido em (3.5), tenha variância não-negativa.

Além disto, a função de correlação normalmente necessita possuir as seguintes características (Diggle & Ribeiro, 2000):

a) $\rho(\cdot)$ deve ser uma função monótona não decrescente em h. Esta restrição tem origem nas idéias fundamentais da dependência espacial, ou seja, que a correlação entre dois pontos diminui à medida que a distância entre eles aumenta;

b) $\rho(h) \to 0$ quando $h \to \infty$, ou seja, a correlação entre dois pontos muito distantes é nula;

c) A função $\rho(h)$ necessita possuir pelo menos um parâmetro que controle a "taxa de decréscimo" no qual a correlação tende à zero.

A função de correlação ou correlograma $\rho(h)$ normalmente possui um ou dois parâmetros, enquanto que o variograma possui três ou quatro — os dois parâmetros pertencentes ao correlograma, o valor de σ^2 e o valor da pepita. Vamos agora analisar algumas famílias de correlação amplamente utilizadas.

3.6.1. FAMÍLIA ESFÉRICA

A família esférica de funções de correlação possui apenas um parâmetro e é definida por

$$\rho(h;\phi)=\begin{cases}1-\frac{3}{2}(h/\phi)+\frac{1}{2}(h/\phi)^3 : 0\le h\le\phi \\ 0 \qquad\qquad\qquad\qquad\quad : \mathrm{h}\ge\phi\end{cases} \tag{3.8}$$

onde ϕ é o parâmetro que controla a taxa de decréscimo da função de correlação. Quando trabalhamos com uma função de correlação da família esférica em Geoestatística, este parâmetro define a amplitude ou alcance da dependência espacial, pois somos "obrigados" a aceitar que a correlação entre dois pontos localizados à uma distancia $\mathrm{h}>\phi$, é *nula*.

A função de correlação esférica $\rho(h;\phi)$ representada na Figura 3.6 é contínua e duplamente diferenciável na origem e, portanto, segundo o teorema 3 ela é a função de correlação correspondente a um processo $S(x)$ diferenciável em média quadrática (Diggle & Ribeiro, 2000).

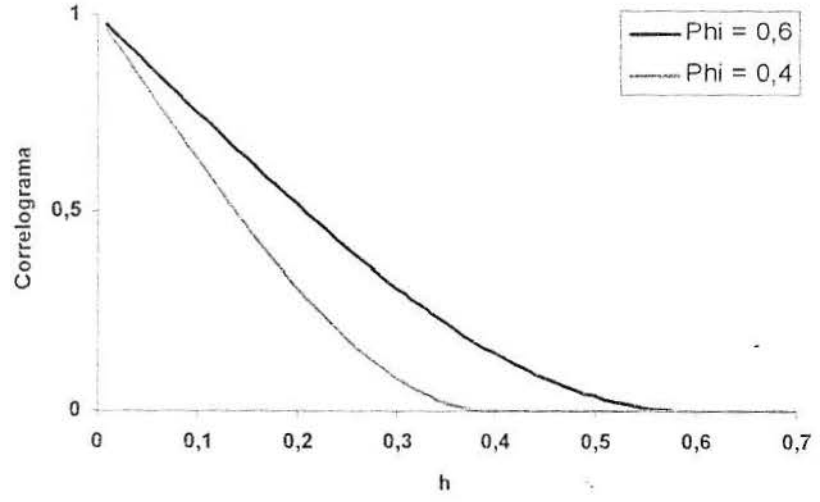

Figura 3.6 - Função de Correlação Esférica para $\phi = 0,4$ e $\phi = 0,6$

3.6.2. FAMÍLIA EXPONENCIAL POTÊNCIA

Esta família de funções de correlação exponencial potência possui dois parâmetros e é definida por

$$\rho(h;\phi;k)=\exp\{-(h/\phi)^2\} : \phi>0 \quad e \quad 0<k\le 2\,. \tag{3.9}$$

O processo subjacente $S(x)$ correspondente a esta família de funções de correlação é contínuo em média quadrática se $k < 2$, mas não é diferenciável nesta circunstância. Entretanto, quando $k = 2$, o processo $S(x)$ torna-se infinitamente diferenciável.

Quando o valor de k é igual à 1 (um), esta função se torna conhecida por *função de correlação exponencial*, a qual é amplamente utilizada em Geoestatística. Quando o valor de k é igual à 2 , temos a *função de correlação Gaussiana*.

Diferentemente da família esférica, esta família de funções não possui uma amplitude, ou seja, a correlação espacial é maior do que zero para $h > \phi$. Isto faz com que a correlação entre dois pontos extremamente distantes seja não-nula, embora pequena.

Segundo Diggle & Ribeiro (2000), esta família de funções frequentemente fornece um ajuste razoável para a estrutura de correlação espacial dos dados. Entretanto, predições baseadas nesta família de funções normalmente tendem a não ser robustas em relação a pequenos desvios do modelo assumido.

Outra desvantagem destacada, é a de que esta família de funções não é muito flexível no que diz respeito à sua forma. Estas funções mudam repentinamente de comportamento quando passamos de um valor de $k < 2$ para um valor de $k = 2$.

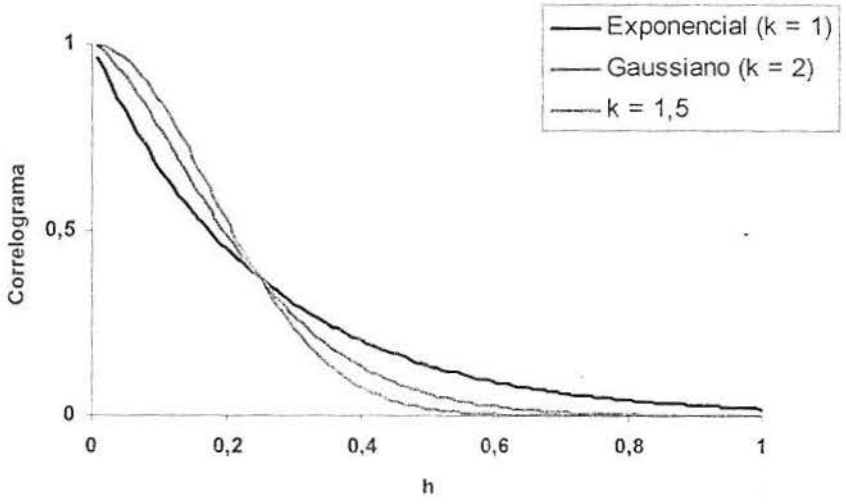

Figura 3.7 - Função de Correlação Exponencial Potência para $\phi = 0,25$ e diferentes valores de k

Em muitos livros de Geoestatística, o parâmetro ϕ é tratado como o alcance ou amplitude da dependência espacial do fenômeno em estudo. Após analisar a Figura 3.7, vemos claramente que este tipo de conclusão deve ser evitada. Apesar das três funções apresentarem $\phi = 0,25$, vemos que a correlação espacial para $h = 0,25$ é $\rho(0,25) \cong 0,36$, o qual representa um valor demasiado alto para considerarmos como limite para a influência da dependência espacial entre dois pontos. Considerar o parâmetro ϕ como

uma medida do alcance ou amplitude da dependência espacial é válido somente para alguns modelos de funções de correlação, como é o caso da família de funções esféricas.

3.6.3. FAMÍLIA MATÉRN

A família de funções de correlação Matérn possuem dois parâmetros e são definidas por

$$\rho(h;\phi;k) = \{2^{k-1}\Gamma(k)\}^{-1}(h/\phi)^k K_k(h/\phi): \quad \phi > 0 \quad e \quad k > 0, \qquad (3.10)$$

onde $K_k(\cdot)$ denota a *função Bessel*[8] *modificada de terceiro tipo de ordem k*.

Esta família de funções é equivalente à *função de correlação exponencial* quando temos $k = 0{,}5$ e é semelhante a *função de correlação Gaussiana* quando $k \to \infty$.

A família de funções Matérn (ver Figura 3.8) possui uma vantagem em relação às famílias apresentadas anteriormente que a torna extremamente importante. Esta vantagem está no fato do parâmetro k controlar a diferenciabilidade do processo $S(x)$ subjacente de maneira direta. *A parte inteira de k nos dá o número de vezes que o processo S(x) é diferenciável em média quadrática.* Por exemplo, se o valor de k é igual à 1,5 , significa que o processo subjacente $S(x)$ é diferenciável. Já um valor de k igual à 0,5 , significaria que o processo não é diferenciável, mas apenas contínuo (Diggle & Ribeiro, 2000).

Pelo fato desta família de funções ser flexível e por possuir apenas dois parâmetros, Diggle & Ribeiro (2000) a consideram como a melhor função de correlação para o uso em Geoestatística.

---

[8] As funções Bessel são as soluções para a equação diferencial de Bessel (matemático que viveu de 1784 à 1846) dada por $x^2y'' + xy' + (x^2 - n^2)y = 0$. A equação diferencial possui duas soluções lineares independentes conhecidas por Função de Bessel de primeiro tipo e Função Bessel de segundo tipo. A função Bessel de terceiro tipo é uma combinação complexa entre as duas soluções para a equação diferencial de Bessel, onde a parte real é composta pela função de Bessel de primeiro tipo e a parte complexa é composta pela função Bessel de segundo tipo (Press et. al., 1986).

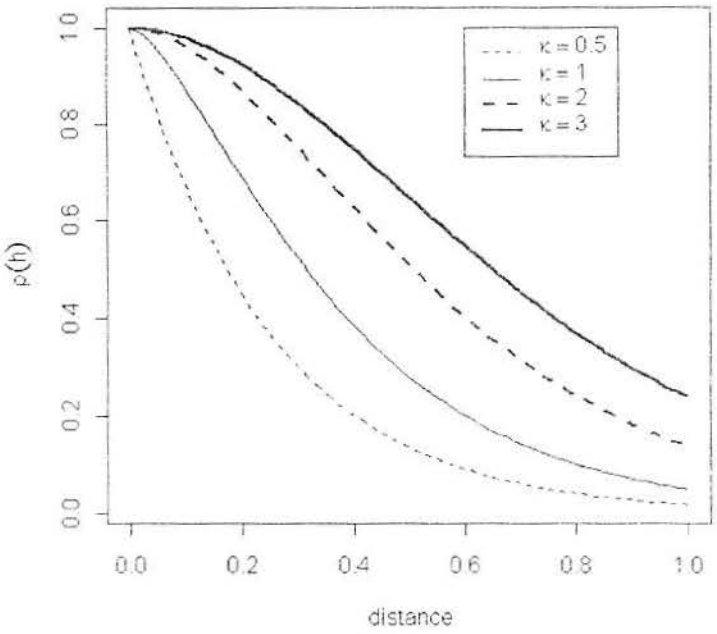

Figura 3.8 - Função de correlação Matérn para diferentes valores de k (ϕ =0.25)

4. ESTIMAÇÃO DA ESTRUTURA DE COVARIÂNCIA ESPACIAL

Neste capítulo apresentaremos maneiras de realizar a escolha de um modelo adequado para a estrutura de covariância espacial, bem como maneiras de estimar os parâmetros envolvidos no modelo escolhido. Apresentaremos também algumas técnicas exploratórias a fim de verificar se as suposições envolvidas no modelo geoestatístico estão satisfeitas. Por fim, realizaremos a estimação da estrutura de covariância espacial utilizando simulações e os dados fornecidos pelo projeto MAPEM através da máxima verossimilhança em comparação com a estimação realizada por mínimos quadrados.

A fim de encontrarmos um modelo de covariância espacial que se ajuste ao fenômeno que estudamos, é comum utilizarmos um estimador do variograma e, a partir disto, escolhermos o modelo apropriado. O motivo da escolha do variograma como ferramenta básica para este processo de estimação está no fato de ele possuir algumas vantagens em relação ao covariograma. Entre estas vantagens destacam-se as seguintes:

a) O variograma é definido para casos onde o covariograma não existe;
b) O vício causado quando nosso processo é estacionário de segunda-ordem é maior para o covariograma do que para o variograma;
c) O variograma não requer estimação da média μ do processo;
d) Nos casos em que não conseguimos detectar que o nosso processo não possui média constante, a estimação do covariograma é afetada numa proporção muito maior do que a estimação do variograma.

Comparações detalhadas entre o desempenho do variograma e do covariograma no processo de estimação da estrutura de covariância espacial são encontrados em Cressie (1991).

4.1. ESTIMAÇÃO DO VARIOGRAMA

Supondo um processo Gaussiano $Y(x)$, como definido no capítulo anterior, temos que o variograma é definido por:

$$\gamma(x - x') = \tfrac{1}{2}\{Var[Y(x) - Y(x')]\}.$$

Adicionalmente, se assumimos que $Y(x)$ é estacionário e isotrópico, temos:

$$\gamma(x-x') = \tfrac{1}{2}\{Var[Y(x)-Y(x')]\}$$
$$\gamma(x-x') = \tfrac{1}{2}\{E[Y(x)-Y(x')-E(Y(x)-Y(x')]^2\}$$
$$\gamma(x-x') = \tfrac{1}{2}\{E[Y(x)-Y(x')-\mu+\mu]^2\}$$
$$\gamma(x-x') = \tfrac{1}{2}\{E[Y(x)-Y(x')]^2\}$$
$$\gamma(h) = \tfrac{1}{2}\{E[Y(x)-Y(x+h)]^2\}, \qquad (4.1)$$

onde h representa a distância euclidiana entre x e x' (Diggle & Ribeiro, 2000).

A partir disto, podemos definir um estimador natural para o variograma, com base em uma amostra $y_1, \ldots y_n$, da seguinte forma:

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{|x_i - x_j| = h} [y(x_i) - y(x_j)]^2, \qquad (4.2)$$

onde o somatório é com relação a todos os pares de observações separadas por uma distância h e $n(h)$ é o número total destes pares (Baley & Gatrell, 1995). Este estimador é normalmente conhecido como *estimador clássico do variograma* (ver Cressie, 1991).

De forma equivalente, o *estimador clássico* do covariograma do processo $Y(x)$ é dado pela seguinte expressão:

$$\hat{C}(h) = \frac{1}{n(h)} \sum_{|x_i - x_j| = h} \{[y(x_i) - \bar{y}(x)][y(x_j) - \bar{y}(x)]\}, \qquad (4.3)$$

onde o somatório é com relação a todos os pares de observações separadas por uma distância h e $n(h)$ é o número total destes pares (Baley & Gatrell, 1995).

Se o processo $Y(x)$ não fosse isotrópico, o número de pares envolvidos nos somatórios dados por (4.2) e (4.3) seriam restritos à direção do vetor h, ou seja, o variograma e o covariograma teriam expressões diferentes dependendo da direção considerada. Quando supomos que a variabilidade de $Y(x)$ entre dois locais x depende adicionalmente da *direção* da distância, chamamos o processo de *anisotrópico*. Normalmente, as direções consideradas são as de $0°$, $45°$, $90°$ e $135°$ (ver Figura 4.1).

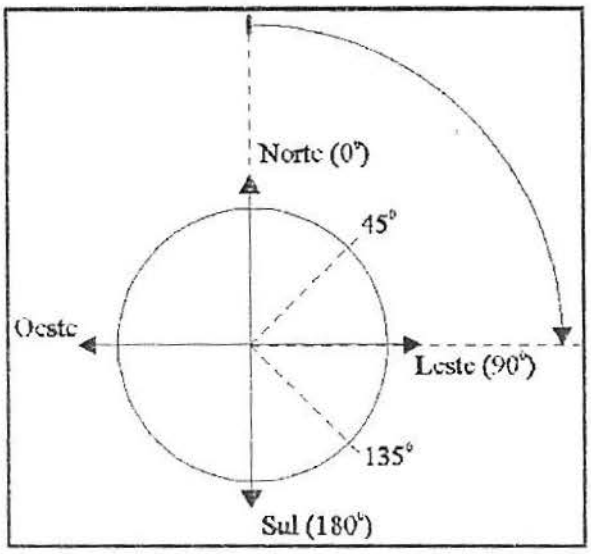

Figura 4.1 - Direções normalmente consideradas em processos anisotrópicos (indicadas pelas linha pontilhadas)[9].

Antes de prosseguirmos o assunto de estimadores para o variograma, vamos ilustrar o processo de estimação do variograma considerando a organização da amostra no espaço.

4.2. MALHA AMOSTRAL

Quando realizamos um delineamento amostral objetivando utilizar técnicas de Geoestatística, geralmente projetamos nossa malha amostral de duas maneiras: *amostras regularmente espaçadas* e *amostras irregularmente espaçadas* (ver Figura 4.5).

[9] Fonte: Eduardo Celso Gerbi Camargo - *Análise Espacial de Superfícies por Geoestatística* — capítulo integrante do livro on-line Análise Espacial de Dados Geográficos (Suzana Fuks, Gilberto Câmara, Antônio M. Monteiro). São José dos Campos, INPE, 2001 (2a. edição, revista e ampliada). http://www.dpi.inpe.br/gilberto/livro/analise/index.html

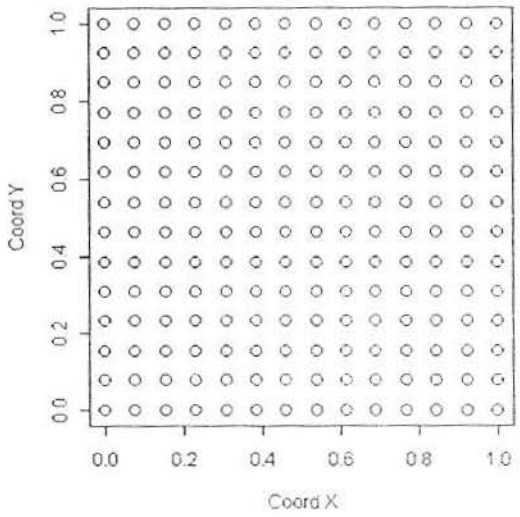

Figura 4.2 - Exemplo simulado de 196 amostras regularmente espaçadas em $\Re^2$.

Quando temos amostras regularmente espaçadas, nosso passo (*lag*) ou distância h entre duas amostras consecutivas é constante.

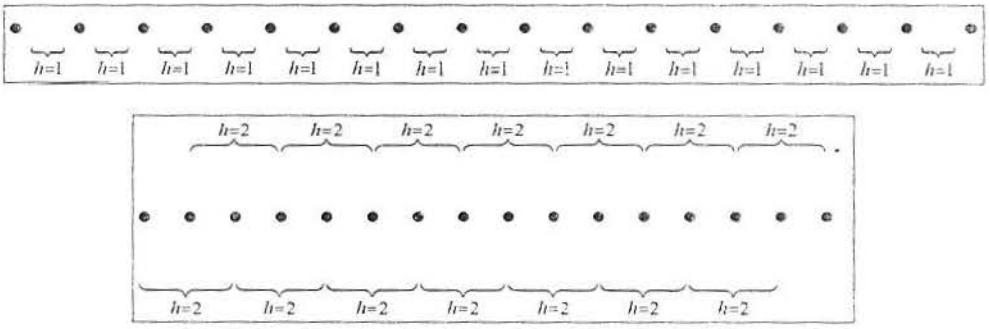

Figura 4.3 - Pares de amostras considerados no cálculo de $\hat{\gamma}(h)$ *para* $h = 1$ *e* $h = 2$ *para o caso de isotropia e amostras regularmente espaçadas em* $\Re^1$.

Para estimarmos o variograma através da expressão (4.2), basta somar as diferenças ao quadrado para todos os pares separados pela distância h. É importante salientar que serão incluídos no cálculo de $\hat{\gamma}(h)$ todos os pares separados pela distância h para *todas as direções conjuntamente*.

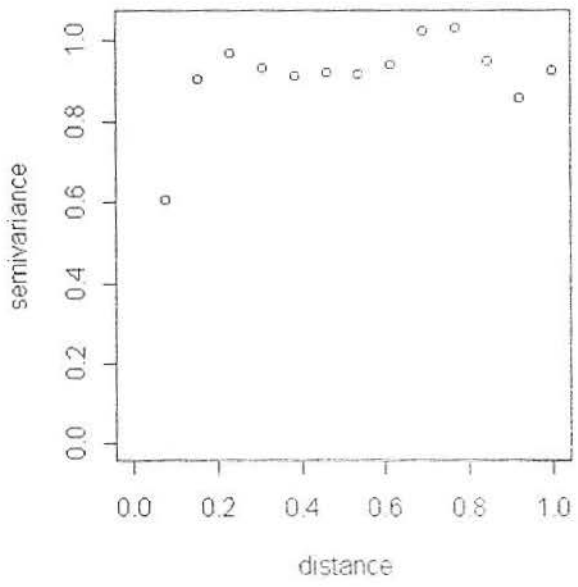

Figura 4.4 - Variograma estimado pela expressão (4.2) a partir de amostras de um processo $S(x)$ simulado na malha regular descrita pela Figura (4.2), cujo modelo de correlação espacial é Esférico com $\tau^2 = 0, \sigma^2 = 1$ e $\phi = 0{,}25$.

Caso nosso processo seja anisotrópico, devemos calcular o valor de $\hat{\gamma}(h)$ para cada direção que julgarmos pertinente. Posteriormente, serão apresentadas maneiras de verificar para quais direções é pertinente calcular $\hat{\gamma}(h)$ ao lidarmos com um processo anisotrópico.

Quando a distância entre duas amostras consecutivas é variável, temos o caso de amostras irregularmente espaçadas (ver Figura 4.5), onde não podemos utilizar as expressões (4.2) e (4.3) diretamente.

Neste caso, devemos especificar *tolerâncias* para as distâncias h e para a direção especificada (caso o processo seja anisotrópico) para procedermos ao cálculo do estimador do variograma.

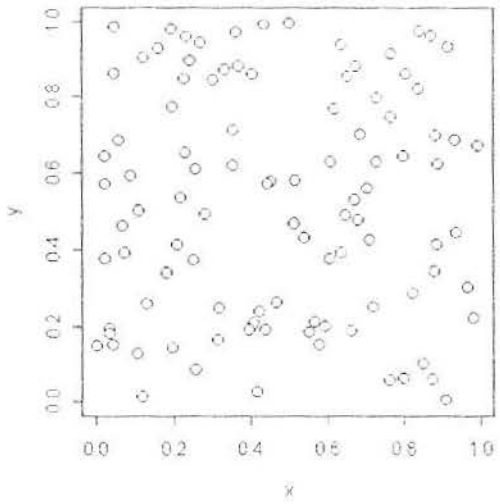

Figura 4.5 - Exemplo simulado de 100 amostras irregularmente espaçadas em $\Re^2$.

A *tolerância* de *lag* pode ser melhor entendida através de um exemplo: Suponha que desejamos calcular $\hat{\gamma}(h)$ para $h = 1\ m$. Se nossos locais x_i da amostra não estiverem regularmente espaçados, dificilmente teremos algum par separado exatamente por um vetor distância $h = 1\ m$. Provavelmente, teremos amostras espaçadas por $0{,}9\ m$, $1{,}1\ m$, $1{,}05\ m$, etc. Se seguíssemos a forma da expressão (4.2) "à risca", poderia acontecer o fato de que nenhum par de pontos pudesse ser incluído no somatório, pois nenhum deles iria satisfazer a exigência de estarem espaçados exatamente $h = 1\ m$. Para resolvermos este problema, poderíamos definir uma tolerância para h. Se a tolerância utilizada fosse de $0{,}1\ m$, todas os pares de amostras separados por uma distância $0{,}9\ m \leq h \leq 1{,}1\ m$ seriam incluídos no cálculo de $\hat{\gamma}(h)$ para $h = 1\ m$.

A *tolerância para a direção* ou *tolerância angular* pode ser entendida de forma semelhante à tolerância de *lag*. Se estamos trabalhando com amostras irregularmente espaçadas e supomos que o processo é anisotrópico — possuindo variogramas diferenciados para as direções de 0°, 45° e 90° —, podemos definir tolerâncias angulares a fim de evitarmos que poucos (ou nenhum) pares de amostras possam ser incluídos no cálculo de $\hat{\gamma}(h)$ para uma dada direção. A Figura 4.4 mostra o funcionamento das tolerâncias de *lag* e *angulares*.

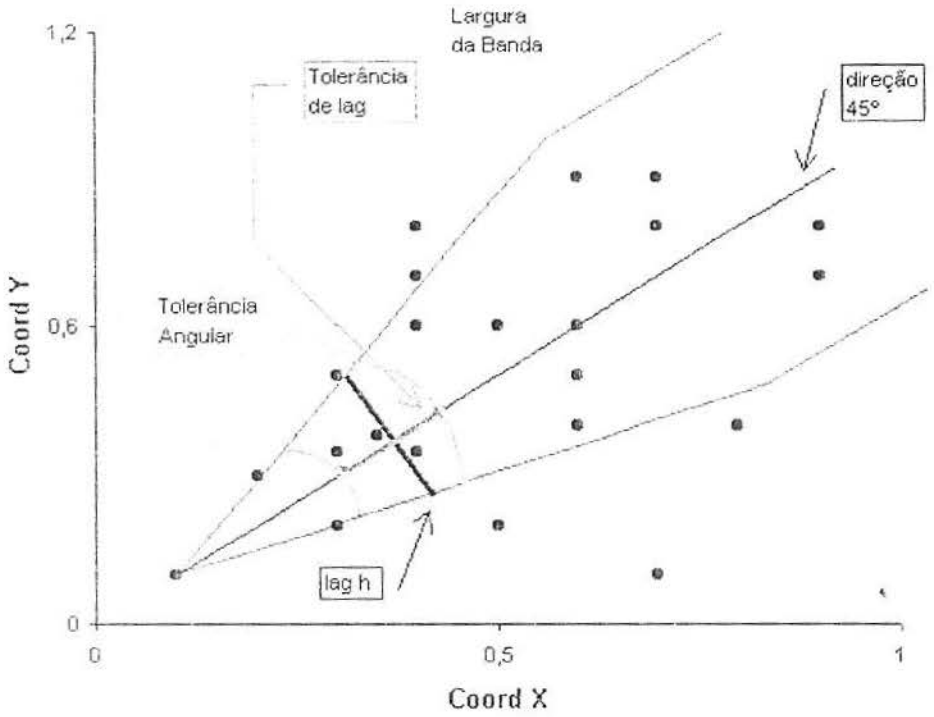

Figura 4.6 - Tolerâncias para um determinado lag h e direção 45° a partir de amostras irregularmente espaçadas em $\Re^2$.

A largura de banda mostrada na Figura 4.6 se refere a um valor de ajuste a partir do qual se restringe o número de pares de observações para o cálculo do variograma.

4. 3. O VARIOGRAMA EMPÍRICO

Supondo um processo $Y(x)$ estacionário e isotrópico, podemos definir o variograma empírico.

Definição 4.1: O *variograma empírico* de um conjunto de dados $(Y_i, x_i): i = 1, \ldots, n$ é o conjunto de pontos $(h_{ij}, v_{ij}): j > i$, onde $h_{ij} = \|x_i - x_j\|$ e $v_{ij} = \frac{1}{2}(Y_i - Y_j)^2$. Um diagrama de dispersão dos pontos (h_{ij}, v_{ij}) é chamado de *variograma nuvem* (Diggle & Ribeiro 2000).

Este diagrama recebe este nome em virtude da enorme quantidade de pontos que nele são plotados. Para uma amostra de n locais, o variograma *nuvem* terá $\frac{1}{2}n(n-1)$ valores de ordenadas.

Segue de (4.1) que cada v_{ij} é um estimador não-viciado para o correspondente $\gamma(h_{ij})$. Adicionalmente, se $Y(x)$ é Gaussiano, a distribuição amostral de cada v_{ij} é proporcional a $\chi^2_{(1)}$. Logo,

$$v_{ij} \sim \gamma(h_{ij})\chi^2_{(1)}$$

$$E(v_{ij}) = \gamma(h_{ij})$$

$$Var(v_{ij}) = 2\gamma(h_{ij})^2$$

Como existe uma variabilidade muito grande no variograma *nuvem* (ver Figura 4.2), é difícil identificar um modelo para $\gamma(h)$. É importante notar também que a informação sobre $\gamma(h)$ para pequenos valores de h é muito maior do que para grandes valores de h.

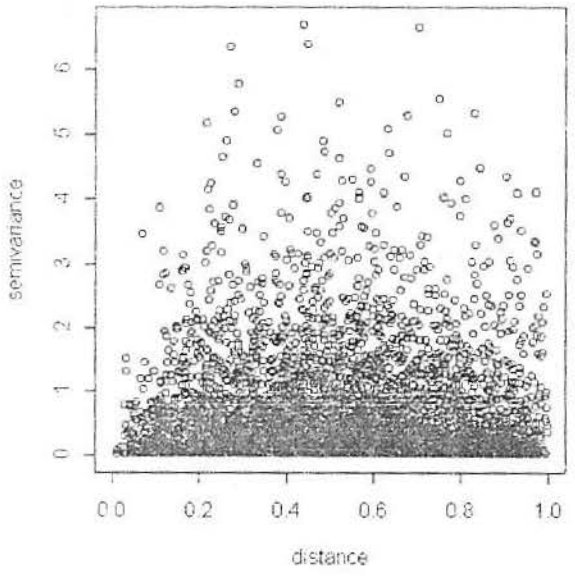

Figura 4.7 - Variograma nuvem estimado a partir de amostras de um processo $S(x)$ simulado na malha irregular descrita pela Figura (4.5), cujo modelo de correlação espacial é Esférico com $\tau^2 = 0, \sigma^2 = 1$ e $\phi = 0{,}25$.

4.4. SUAVIZAÇÃO DO VARIOGRAMA EMPÍRICO

Quando possuímos amostras regularmente espaçadas, a expressão do estimador do variograma dada por (4.2) é obtida de forma direta. Ao fazer o variograma *nuvem*, obteríamos um gráfico semelhante ao mostrado na Figura 4.8.

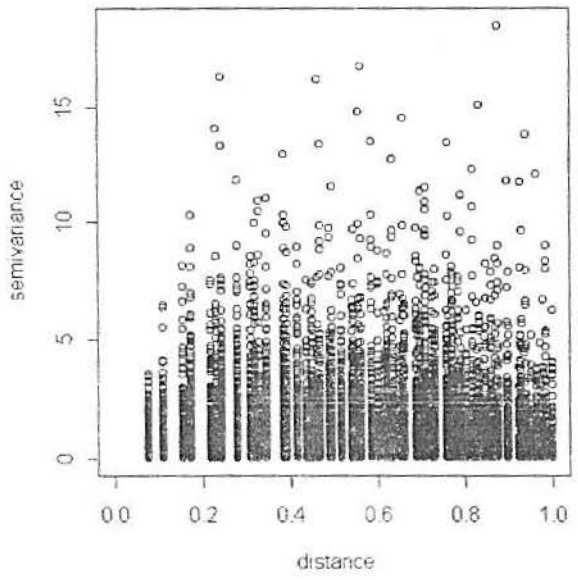

Figura 4.8- Variograma nuvem estimado a partir de amostras de um processo $S(x)$, simulado na malha regular descrita pela Figura (4.2), cujo modelo de correlação espacial é Esférico com $\tau^2 = 0, \sigma^2 = 1$ e $\phi = 0{,}25$.

A fim de suavizar este gráfico e resumir a informação, podemos tomar a média dos v_{ij} para cada *lag* ou distância h. Procedendo assim, obteríamos um gráfico como mostra a Figura 4.9.

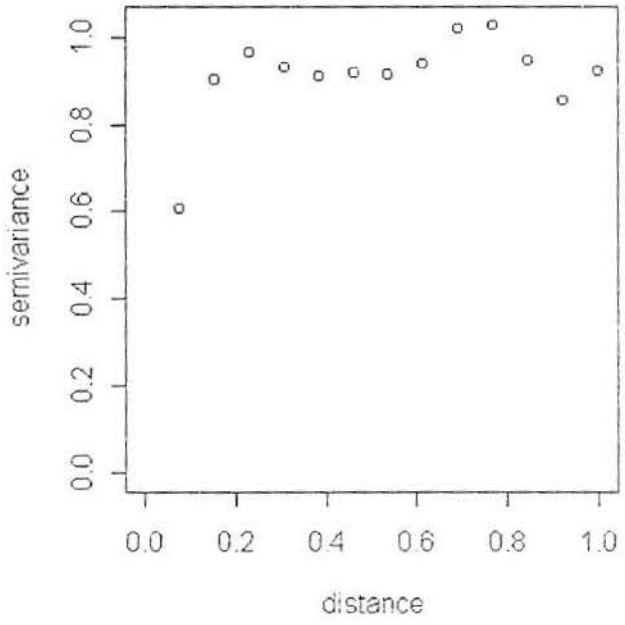

Figura 4.9 - Variograma suavizado do processo $S(x)$ simulado na malha regular descrita pela Figura 4.2, cujo modelo de correlação espacial é Esférico com $\tau^2 = 0, \sigma^2 = 1$ e $\phi = 0{,}25$.

Como podemos notar, esta suavização é equivalente ao estimador clássico do variograma dado pela expressão (4.2).

Entretanto, quando nossas amostras são irregularmente espaçadas, o variograma *nuvem* é mais complexo. Neste caso, para suavizá-lo é necessário utilizar idéias semelhantes aos conceitos de tolerâncias citados anteriormente.

Definição 4.2: O *variograma amostral* é o conjunto de pontos $(h_k, \bar{v}_k)$, onde $h_k : k = 1, \ldots, m$ é um conjunto predeterminado de distâncias, $\bar{v}_k = \frac{1}{n_k} \sum_{h_{ij} \in S_k} v_{ij}$, S_k é um conjunto de valores de h mais próximos de h_k do que qualquer outro $h_{k'}$, e n_k é o número de h_{ij} em S_k (Diggle & Ribeiro, 2000).

A partir da definição 4.2, segue que o estimador dado pela expressão (4.2) é o *variograma amostral* supondo amostras regularmente espaçadas e isotropia.

Podemos notar também que quando lidamos com amostras irregularmente espaçadas, dependendo do número de conjuntos de distâncias k escolhido, temos um variograma amostral diferente. A Figura 4.10 ilustra este fato:

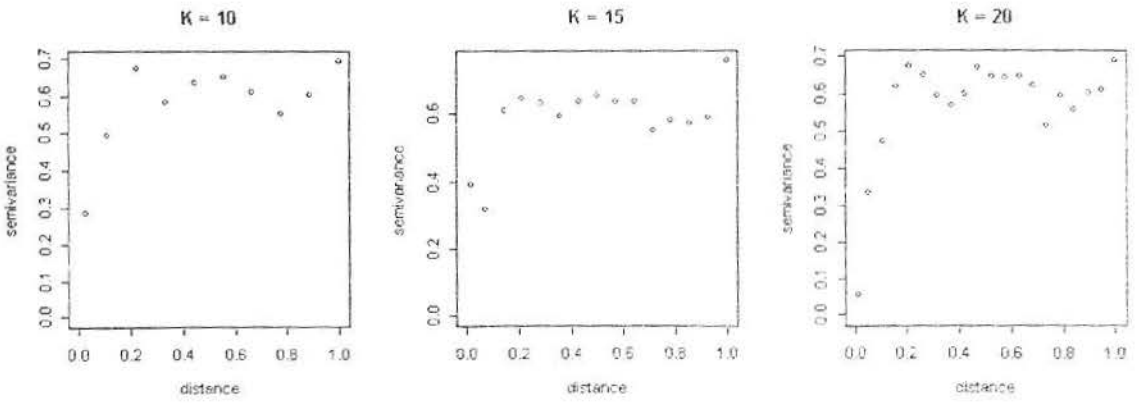

Figura 4.10 - Variogramas amostrais do processo $S(x)$ simulado na malha irregular descrita pela Figura 4.5, cujo modelo de correlação espacial é Esférico com $\tau^2 = 0, \sigma^2 = 1$ e $\phi = 0{,}25$, calculados para 3 valores de k diferentes.

A partir do cálculo do variograma amostral, resumimos a informação do variograma empírico. Entretanto, precisamos que o variograma amostral seja contínuo para que possamos realizar predições para locais x separados por qualquer distância h.

Partindo para uma segunda etapa do processo de suavização, podemos ajustar uma linha aos variogramas amostrais, como mostra a Figura 4.11:

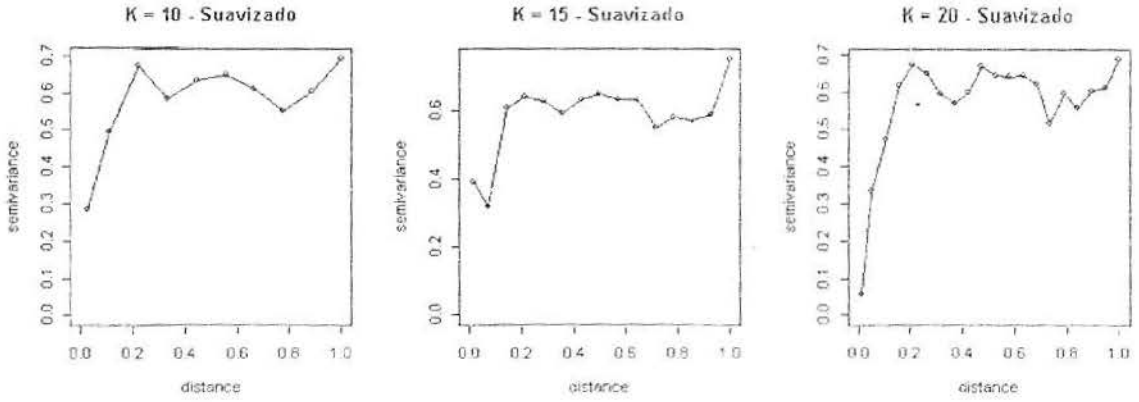

Figura 4.11 - Segunda etapa de suavização dos variogramas amostrais plotados na Figura 4.10.

Outro método de suavização do variograma empírico é o *suavizador Kernel*. Este método define o estimador do variograma utilizando ponderações baseadas em uma função Kernel e uma amplitude de banda. Para maiores considerações ver Diggle & Ribeiro (2000).

Apesar da utilidade da suavização do variograma empírico, um variograma *nuvem* pode revelar distorções e valores extremos que dominam a estimativa do variograma amostral. Ele pode também revelar, através de uma análise destas distorções

para qualquer *lag h*, que o variograma amostral será uma pobre estimativa da verdadeira estrutura de covariância espacial (Bailey & Gatrell,1995).

4.5. EFEITOS DE TENDÊNCIA NO PROCESSO DE ESTIMAÇÃO DO VARIOGRAMA

Quando não detectamos que o nosso processo $Y(x)$ não possui a média constante, a informação embutida no variograma empírico torna-se altamente duvidosa. Uma maneira visual de verificar se o processo possui média constante é fazer um gráfico de superfície em 3-D e buscar observar o comportamento dos valores de $Y(x)$.

Entretanto, nos deparamos com um dilema ao analisar esta superfície: as variações globais (aumentos e decréscimos) dos valores de $Y(x)$ que observamos são causadas por uma *tendência determinística* ou são efeitos da *estrutura de covariância espacial*?

De uma maneira mais formal, se nossos dados seguem o modelo

$$Y_i = \mu(x_i) + S(x_i) + Z_i : i = 1,\ldots,n,$$

onde $\mu(x_i)$ é uma função de uma ou mais variáveis explicativas espaciais (normalmente função dos eixos de coordenadas), então não podemos distinguir entre a função determinística $\mu(x_i)$ e a realização do processo estocástico $S(x)$ sem realizarmos suposições paramétricas específicas. Somente poderíamos fazer esta distinção se pudéssemos replicar o processo $S(x)$ e obter outro conjunto de dados com a mesma função $\mu(\cdot)$.

Uma solução usual é assumir que $\mu(x)$ pode ser descrita por um modelo de regressão usando as variáveis explicativas que referenciam nossos dados no espaço. Em outras palavras, nós inicialmente realizamos uma *análise de superfícies de tendência* ajustando uma função das coordenadas dos eixos (normalmente latitude e longitude) por mínimos quadrados ordinários. Após o ajuste, tomaríamos os resíduos deste ajuste em lugar dos nossos dados Y_i. Ou seja,

$$Y_i^* = Y_i - \hat{\mu}(x_i).$$

A partir dos resíduos Y_i^* podemos proceder à estimação do variograma e da estrutura de covariância espacial. Após realizado o processo de predição espacial, onde

obtemos uma superfície estimada $\hat{S}^{*}(x)$, basta adicionar $\hat{\mu}(x)$ aos valores preditos para obtermos a superfície $\hat{S}(x)$.

Segundo Diggle & Ribeiro (2000), o ajuste de uma função linear é suficiente para remediar o problema da tendência, não sendo necessário a busca por funções mais complexas.

A fim de realizarmos uma análise visual e analisar alguns dos efeitos da presença de tendência no variograma amostral, vamos utilizar uma das variáveis do projeto MAPEM (ver anexos para maiores detalhes sobre o projeto).

Ao teor de *Cromo* nos 47 locais amostrados no estudo, vamos adicionar um componente de tendência linear e calcular o variograma empírico.

Analisando as superfícies estimadas para o teor de *Cromo* (Figura 4.12), vemos que os dados originais não parecem apresentar tendência determinística, enquanto que os dados modificados parecem possuir valores mais altos para o teor de *Cromo* à nordeste da região.

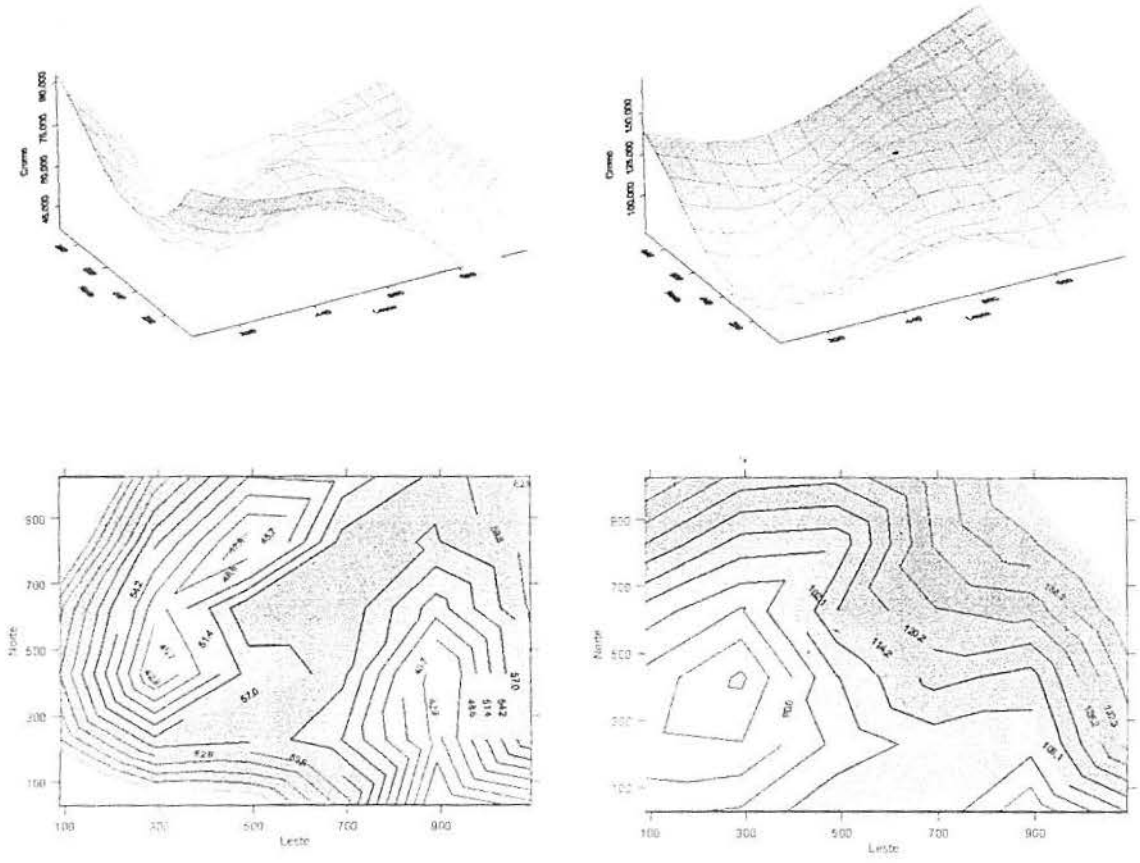

Figura 4.12 - Superfícies e gráficos de contornos ajustados para os teores de Cromo sem tendência (esquerda) e com tendência determinística (direita), para os dados do projeto MAPEM.

Neste caso, a tendência determinística adicionada aos dados originais é uma função linear de valores padronizados das coordenadas, ou seja,

$$\mu_1(x) = 0{,}06z + 0{,}04w \ ,$$

onde (z, w) representam os valores das coordenadas de referência fornecidas pelo GPS, subtraídas de constantes (ver Figura 4.13).

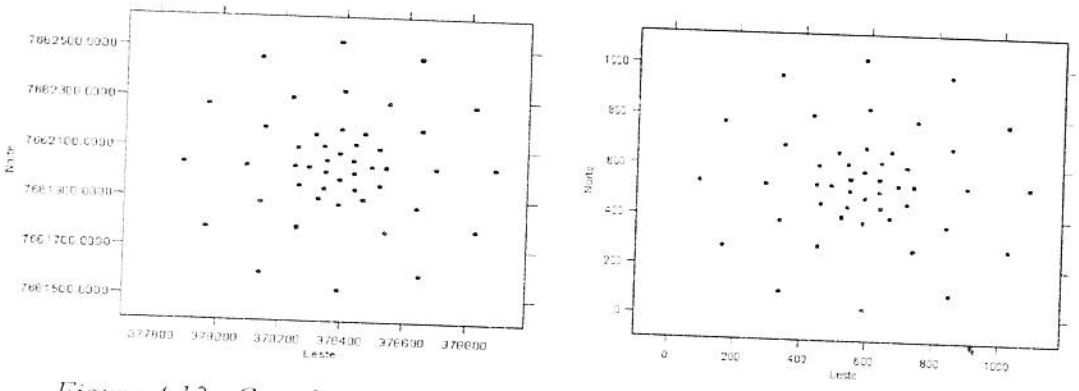

Figura 4.13 - Coordenadas amostrais referenciadas pelo GPS (esquerda) e transformadas para ajuste de tendência (direita).

Calculando o variograma amostral para os teores de *Cromo*, vemos que o variograma correspondente aos dados com tendência parece mais "comportado" que o variograma para os dados originais (Figura 4.14). Isto acontece porque os efeitos de 1ª ordem tendem a ser responsáveis pelo maior percentual da variabilidade na região de estudo.

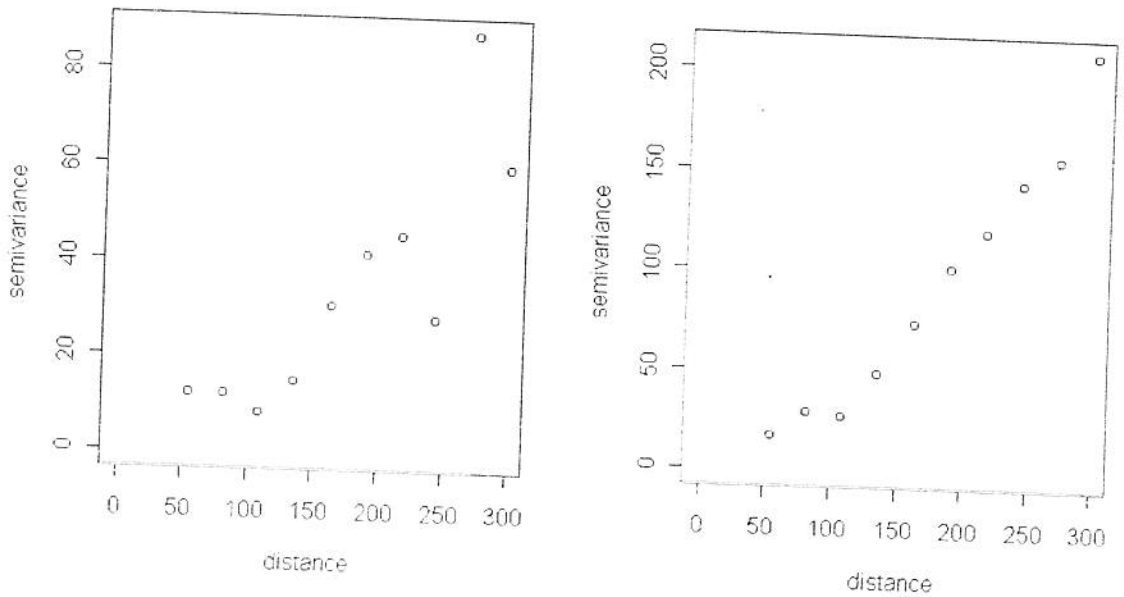

Figura 4.14 - Variogramas amostrais para os teores de Cromo sem tendência (esquerda) e com tendência determinística (direita) para os dados do projeto MAPEM.

Adicionando-se aos dados originais a tendência $\mu_2(x) = 3{,}7z + 1{,}8w$, obteríamos o variograma amostral dado pela Figura 4.15. Se julgássemos que os efeitos de 1ª ordem são constantes, ou seja, que $\mu(x) = \mu$, seríamos levados a crer que a dependência espacial *é ainda melhor definida*. Isto pode ser notado aumentando-se o número de conjuntos de distâncias k no variograma amostral. Mesmo aumentando o valor de k, o comportamento do variograma permanece quase inalterado.

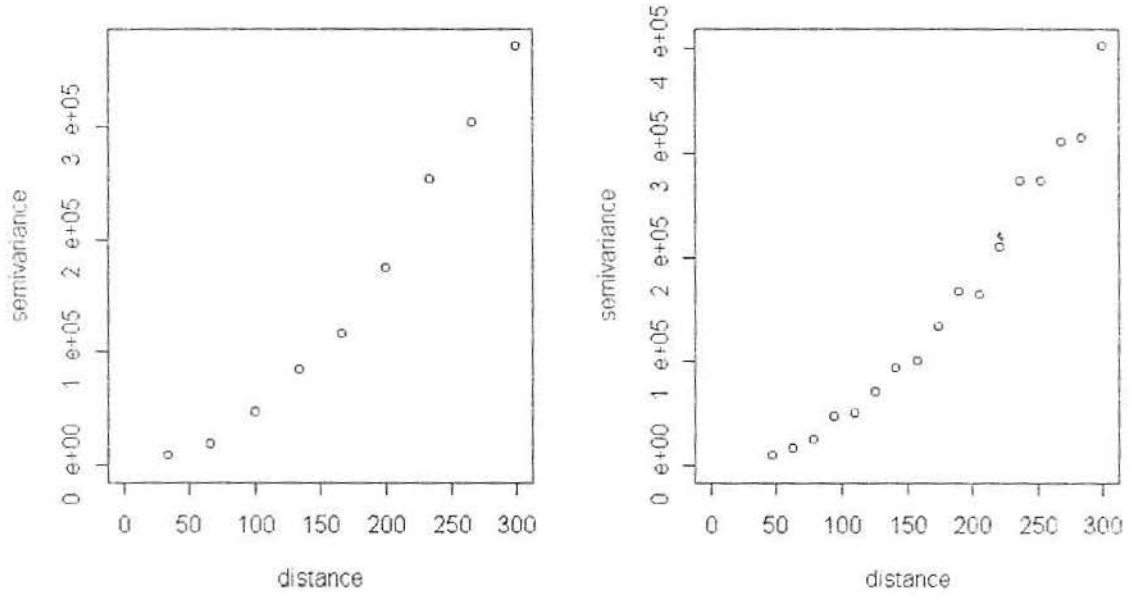

Figura 4.15 - Variograma amostral para os teores de Cromo com tendência determinística dada por $\mu_2(x)$, calculado com k = 10 (esquerda) e k = 18 (direita).

Por este motivo, é de vital importância analisar a região D em estudo antes de realizar as análises geoestatísticas. É preciso uma boa avaliação do fenômeno antes de julgar se os dados possuem média constante ou se possuem uma tendência determinística.

4.6. EFEITOS DE VALORES ATÍPICOS NO PROCESSO DE ESTIMAÇÃO DO VARIOGRAMA

Os *outliers* ou dados atípicos possuem indesejável efeito no processo de estimação do variograma. Em um variograma empírico, sabemos que n observações geram $\frac{1}{2}n(n-1)$ ordenadas para serem plotadas no variograma *nuvem*. Isto quer dizer que a influência de apenas 1 (um) outlier gera efeitos em $n-1$ ordenadas do variograma empírico.

Quando um outlier, cujo valor observado é alto, se localiza próximo de locais onde os valores observados são baixos, uma análise do variograma empírico pode conduzir a conclusões erradas sobre o efeito pepita (τ^2) e/ou a existência de correlação espacial (Diggle & Ribeiro, 2000).

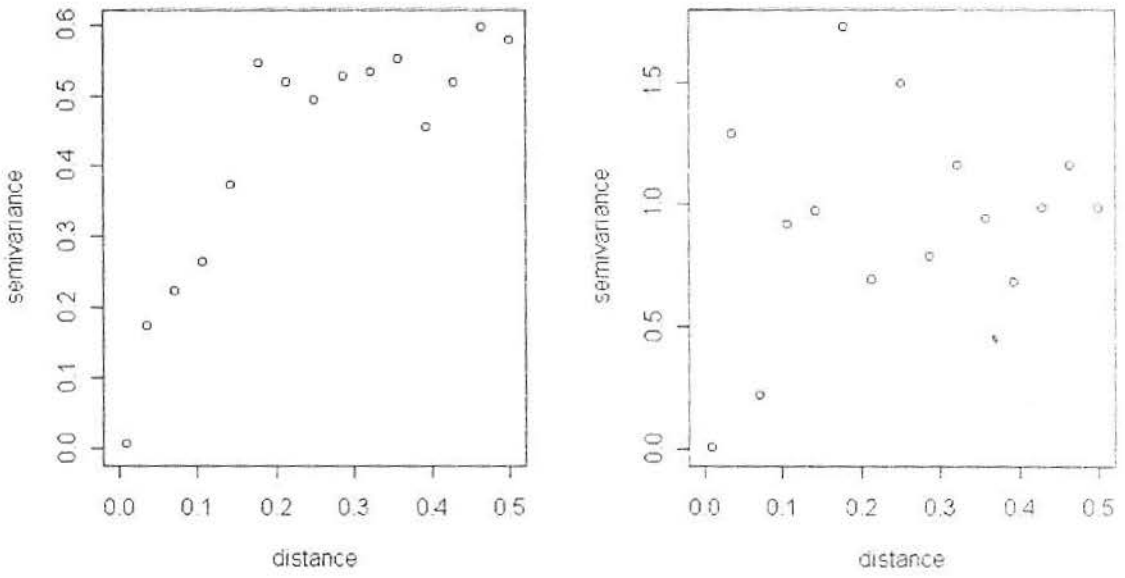

Figura 4.16 - Variogramas amostrais do processo $S(x)$, simulado em malha irregular composta de 100 amostras, cujo modelo de correlação espacial é Esférico com $\tau^2 = 0, \sigma^2 = 1$ e $\phi = 0{,}25$, sem a presença de outliers (esquerda) e com a presença de 1 (um) outlier.

Observando a Figura 4.16 vemos a influência de um outlier no variograma amostral ao modificarmos um valor de "–0,9" unidades para "8" unidades em uma amostra de 100 locais, onde os valores observados do processo estão no intervalo $-0{,}9 \leq y \leq 2{,}52$.

Neste caso, podemos ver claramente a influência desta observação atípica nas demais 99 observações. Sem a presença do outlier, é razoável aceitar que um modelo esférico de correlação espacial é o modelo gerador dos dados, enquanto que com a presença do outlier, não pode-se distinguir nenhum modelo razoável para a correlação espacial.

Com base nestas duas últimas seções, *influência da tendência e da presença de outliers na estimação do variograma*, comprovamos a importância de uma análise preliminar nos dados a fim de tentar detectar qualquer sinal de falta de estacionariedade no processo em estudo. Procedendo desta forma, estaremos nos prevenindo de vícios que comprometam a análise geoestatística dos dados.

4.7. ESTIMADOR ROBUSTO DO VARIOGRAMA

Uma técnica robusta é aquela que permanece "estável" mesmo quando se depara com desvios das suposições assumidas no modelo. No nosso contexto, os desvios podem ser considerados como contaminações no processo Gaussiano $Y(x)$.

Um estimador do variograma, proposto por Cressie e Hawkins, é dado pela seguinte expressão (Cressie, 1991):

$$\bar{\gamma}(h) = \frac{\left\{ \frac{1}{2n(h)} \sum_{|x_i - x_j| = h} \left| y(x_i) - y(x_j) \right|^{1/2} \right\}^4}{(0{,}457 + 0{,}494 / n(h))},$$

onde o somatório é com relação a todos os pares de observações separadas por uma distância h e $n(h)$ é o número total destes pares. O denominador da expressão serve para tornar o estimador não-viciado.

Basicamente, este estimador utiliza a raiz quadrada do módulo das diferenças ao invés de elevá-las ao quadrado. Isto é assim realizado para tornar estas diferenças mais próximas de uma distribuição Gaussiana. Isto decorre do fato de $[y(x_i) - y(x_j)]^2$ ter distribuição $\chi^2_{(1)}$. A raiz quarta é a transformação potência que torna estas diferenças ao quadrado Gaussianas, ou seja, $\left| y(x_i) - y(x_j) \right|^{1/2}$ (Cressie, 1991).

Este estimador $\bar{\gamma}(h)$ é robusto na presença de outliers que contaminam nosso processo e nos casos em que o efeito pepita é alto. Para maiores detalhes consultar Cressie, 1991.

As Figuras 4.17 e 4.18 ilustram a distribuição das ordenadas do variograma empírico estimado pelo método clássico em comparação com as estimativas realizadas pelo estimador robusto.

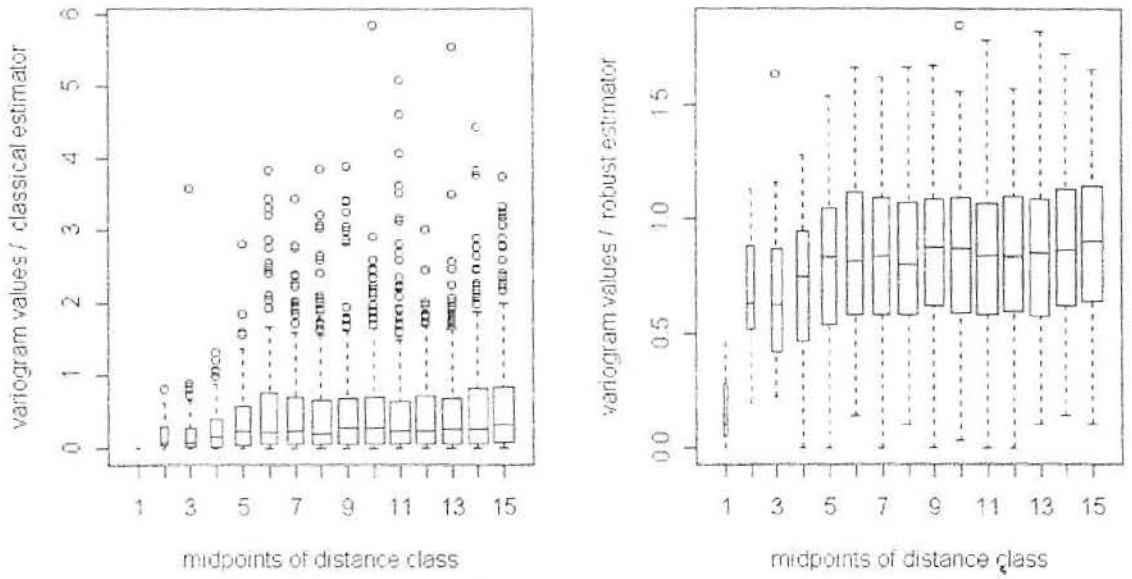

Figura 4.17 - Box-plots mostrando a distribuição das ordenadas do variograma empírico do processo simulado $S(x)$ (apresentado na Figura 4.16), estimadas pelo método clássico (à esquerda) e estimadas pelo método robusto (à direita) sem a presença de outliers.

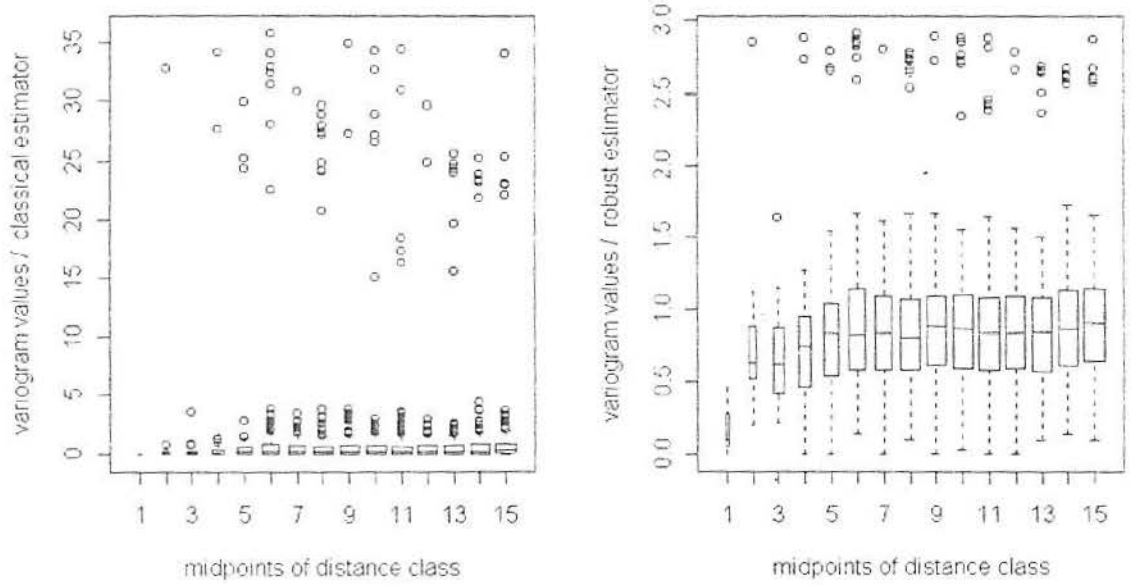

Figura 4.18 - Box-plots mostrando a distribuição das ordenadas do variograma empírico do processo simulado $S(x)$ (apresentado na Figura 4.16), estimadas pelo método clássico (à esquerda) e estimadas pelo método robusto (à direita) com a presença de 1 (um) outlier.

4.8. ESTIMAÇÃO PARAMÉTRICA DA ESTRUTURA DE COVARIÂNCIA ESPACIAL

Os métodos de suavização de variogramas empíricos, descritos nas seções anteriores, não nos dão subsídios para ajustarmos alguma das estruturas de covariância espacial conhecidas (*Matérn*, Esférica, etc.). Tais métodos são conhecidos como estimadores *não-paramétricos* da estrutura de covariância espacial.

Nesta seção, vamos apresentar três métodos de estimação paramétrica de estruturas de covariância espacial: *Mínimos Quadrados Ordinários*, *Mínimos Quadrados Ponderados* e *Máxima Verossimilhança*.

4.8.1. MÍNIMOS QUADRADOS ORDINÁRIOS

Sabemos que as ordenadas v_{ij} do variograma empírico são estimadores não-viciados de $\gamma(h_{ij})$. Levando em conta que um modelo teórico de variograma define uma família de funções $\gamma(h;\theta)$, podemos obter estimadores consistentes de θ no modelo paramétrico $\gamma(h) = \gamma(h;\theta)$ minimizando

$$MQO_1(\theta) = \sum_{j>i} \{v_{ij} - \gamma(h_{ij};\theta)\}^2 \qquad (4.4)$$

ou

$$MQO_2(\theta) = \sum_{j>i} \{\overline{v}_{ij} - \gamma(h_k;\theta)\}^2 \qquad (4.5)$$

que se baseia no variograma amostral.

A expressão dada por (4.5) é normalmente mais utilizada devido à conveniências computacionais. Entretanto, quando trabalhamos com delineamentos de amostras irregularmente espaçadas, as ordenadas v_k são somente aproximadamente não-viciadas para $\gamma(h_k)$. Este vício é conhecido como *vício de suavizamento* (Diggle & Ribeiro, 2000).

A Figura 4.19 mostra o variograma para os teores de areia na região estudada pelo projeto MAPEM, ajustado pelo critério de MQO para duas funções de correlações diferentes.

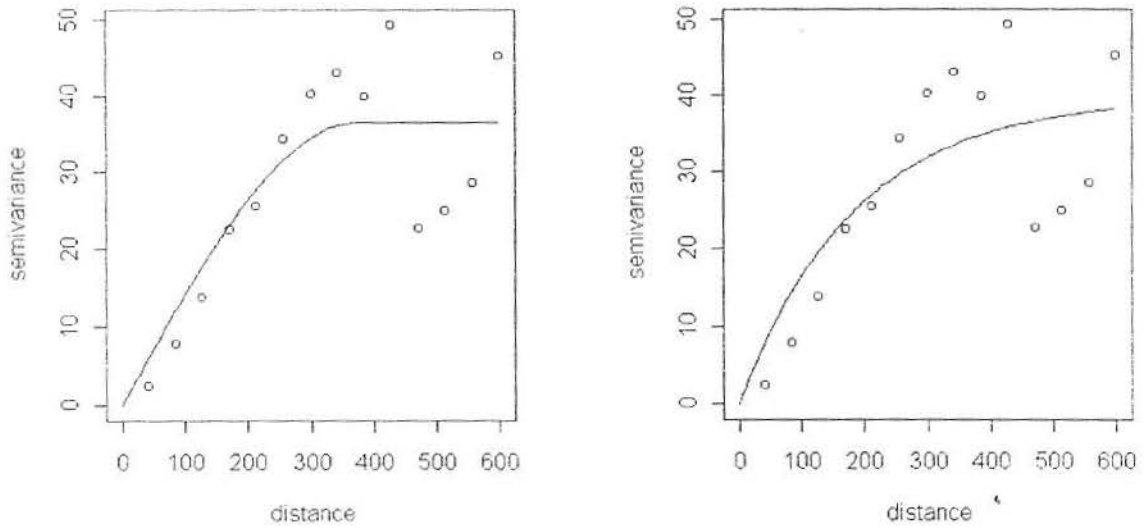

Figura 4.19 - Variograma para o teor de areia na região estudada pelo projeto MAPEM com ajuste pelo critério MQO de uma função de correlação esférica com parâmetros estimados $\hat{\theta} = (\sigma^2 = 36{,}4;\ \phi = 376;\ \tau^2 = 0)$ *(esquerda) e com ajuste de função exponencial com parâmetros estimados* $\hat{\theta} = (\sigma^2 = 39{,}7;\ \phi = 187{,}1;\ \tau^2 = 0)$ *(direita).*

4.8.2. MÍNIMOS QUADRADOS PONDERADOS

Os métodos de estimação por *MQO*, dado por (4.4) e (4.5), possuem ainda outra desvantagem por não considerarem a variabilidade não-constante de v_{ij}. Como vimos anteriormente, supondo um processos Gaussiano, os v_{ij} possuem distribuição $\chi^2_{(1)}$ e variância $2\gamma(h_{ij};\theta)^2$. Este fato, adicionado ao fato de que $\bar{v}_k$ são médias de ordenadas do variograma empírico baseadas em n_k's diferentes, Cressie[10] (1985), apud Diggle & Ribeiro (2000), propôs o seguinte estimador por mínimos quadrados ponderados

$$MQP_2(\theta) = \sum_k n_k \left[\frac{\bar{v}_k - \gamma(h_k;\theta)}{\gamma(h_k;\theta)} \right]^2 \qquad (4.6)$$

Normalmente é recomendado incluir no somatório apenas os $\bar{v}_k$ baseados em $n_k > 30$ (Journel and Huijbregts, 1978). Cressie justificou esta recomendação

[10] Cressie, N.A.C. (1985). *Fitting variogram models by wighted least squares*. Journel of the International Association of Mathematical Geology, 17, 563-86

estabelecendo a consistência deste procedimento assintoticamente quando os n_k tornam-se grandes.

O principal problema que pode ser apontado no critério dado por (4.6) é o fato de que as ponderações que são dadas baseiam-se em parâmetros desconhecidos. Assim, se definimos um critério de minimização baseado nas ordenadas do variograma empírico,

$$MQP_1(\theta) = \sum_{j>i}\left[\frac{v_{ij} - \gamma(h_{ij};\theta)}{\gamma(h_{ij};\theta)}\right]^2 \qquad (4.7)$$

esta desvantagem seria ainda mais evidente.

Outros estudos mostraram que o vício no processo de estimação dado por (4.6) é da ordem de $\frac{1}{n_k}$ (Barry et. al.[11], 1998 apud Diggle & Ribeiro, 2000).

O critério de *MQO* dado pela expressão (4.4) não apresenta este vício. Entretanto, este critério se constitui de um simples mas ineficiente método de estimação dos parâmetros do variograma.

Dentro dos critérios de mínimos quadrados ponderados, define-se ainda um outro critério, baseado na expressão (4.6), que pondera considerando apenas os n_k. Sua expressão é dada por

$$MQP_3(\theta) = \sum_k n_k[\bar{v}_k - \gamma(h_k;\theta)]^2 . \qquad (4.8)$$

Este critério é aproximadamente equivalente ao *MQO*. A diferença entre os dois procedimentos é resultante do vício de suavização ao computar-se os $\bar{v}_k$. Este critério é chamado de *mínimos quadrados n-ponderados* (Diggle & Ribeiro, 2000).

A Figura 4.20 mostra o variograma para os teores de areia na região estudada pelo projeto MAPEM, onde ajustou-se uma mesma função de correlação pelos critérios de MQP_2 e MQP_3.

[11] Barry, J.T., Crowder, M.J. and Diggle, P.J (1997). *Parametric estimation of the variogram*. Lancaster University Technical Report.

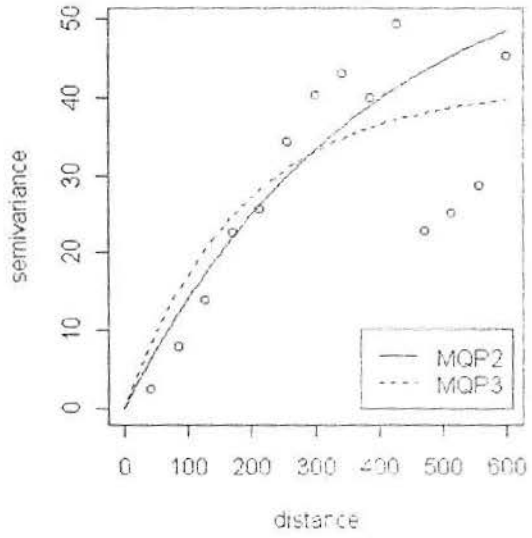

Figura 4.20 - Variograma para o teor de areia na região estudada pelo projeto MAPEM com ajuste pelo critério MQP_2 de uma função de correlação exponencial com parâmetros estimados $\hat{\theta} = (\sigma^2 = 41{,}2;\ \phi = 186{,}5;\ \tau^2 = 0)$ (linha contínua) e com ajuste dado pelo critério MQP_3 com parâmetros estimados $\hat{\theta} = (\sigma^2 = 39{,}7;\ \phi = 187{,}1;\ \tau^2 = 0)$ (linha pontilhada).

Grandes desvantagens dos métodos baseados em ajuste de curvas (MQO e MQP) são notadas quando avaliamos a sensibilidade dos critérios quando variamos as amplitudes h consideradas no processo de estimação. Em outras palavras, se calculássemos o variograma amostral somente para distâncias $h < 300\ m$ e, posteriormente, refizéssemos o cálculo para $h < 350\ m$, obteríamos estimativas dos parâmetros significativamente diferentes. A Figura 4.21 ilustra este fato.

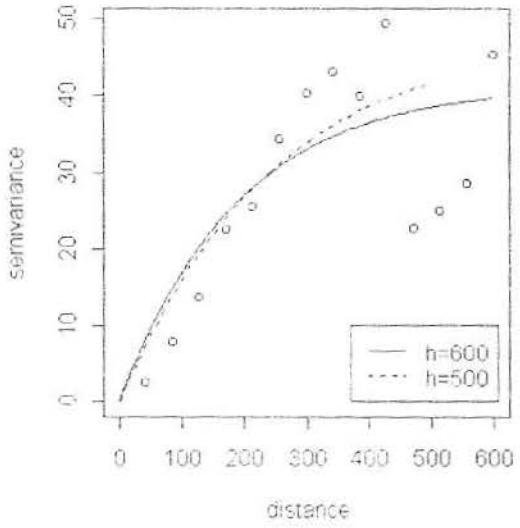

Figura 4.21 - Ajuste de modelo exponencial para o teor de areia pelo critério MQP_3, baseado em variograma amostral com $h = 600$ metros (linha contínua) e com $h = 500$ metros (linha pontilhada).

Por fim, métodos baseados em ajustes de curvas não garantem a legitimidade do modelo estimado. A menos que algumas restrições tenham sido utilizadas no algoritmo computacional de minimização, estes critérios podem produzir estimativas negativas para τ^2 e conduzir a combinações de estimativas que violam a exigência de que as funções de covariância espacial devam ser não-negativas (Diggle & Ribeiro, 2000).

4.8.3. MÉTODOS BASEADOS NA VEROSSIMILHANÇA

Na estatística clássica, estimadores baseados no princípio da máxima verossimilhança tem sido amplamente empregados devido ao fato de geralmente fornecerem estimadores consistentes, assintoticamente eficientes e não-viciados.

Definição 4.3: A estimativa de máxima verossimilhança de θ, isto é, $\hat{\theta}$, baseada em uma amostra aleatória $Y_1, Y_2, \ldots, Y_n$ é aquele valor de θ que torna máxima a função de verossimilhança $L(Y_1, \ldots, Y_n; \theta) = f(Y_1; \theta) f(Y_2; \theta) \ldots f(Y_n; \theta)$ (Meyer, 1983).

Apresentaremos agora o método da máxima verossimilhança para estimação dos parâmetros da estrutura de covariância espacial supondo o modelo Gaussiano como gerador dos dados.

Suponha que os dados $\mathbf{Y} = (Y_1, \ldots, Y_n)$ foram gerados por um modelo Gaussiano linear,

$$Y_i = \mu(x_i) + S(x_i) + Z_i : i = 1, \ldots, n, \tag{4.9}$$

onde $S(x_i)$ é um processo Gaussiano estacionário com variância σ^2, função de correlação $\rho(h;\phi)$ e $Z_i \sim N(0, \tau^2)$. O valor de $\mu(x_i)$ é determinado por um modelo de regressão baseado em k funções de variáveis referenciadas observadas

$$\mu(x_i) = \sum_{k=1}^{p} f_k(x_i)\beta_k . \tag{4.10}$$

O papel da expressão de $\mu(x_i)$ é representar a variabilidade que não é devida à dependência espacial entre os locais x, mas devido a tendência existente na região D.

Apenas para uma melhor compreensão da expressão (4.10), vamos considerar que a regressão é *função quadrática de duas variáveis espacialmente referenciadas* $(z; w)$, representando latitude e longitude, respectivamente. Então podemos escrever $\mu(x)$ da seguinte forma:

$$\mu(x) = \begin{bmatrix} 1 & z_1 & w_1 & z_1^2 & w_1^2 & z_1 w_1 \\ 1 & z_2 & w_2 & z_2^2 & w_2^2 & z_2 w_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_n & w_n & z_n^2 & w_n^2 & z_n w_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix},$$

onde a primeira matriz do lado direito da expressão é chamada de matriz de delineamento. Para maiores detalhes sobre análises de superfícies de tendência, ver Bailey & Gatrell (1995).

A partir das definições dadas por (4.9) e (4.10), segue que

$$\mathbf{Y} \sim MVN(F\beta, G(\theta)),$$

onde

$\theta = (\tau^2, \sigma^2, \phi)$

$$E(\mathbf{Y}) = \mathbf{F}\beta = [\text{matriz de delineamento } n \times p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$Var(\mathbf{Y}) = \mathbf{G}(\theta) = \tau^2\mathbf{I} + \sigma^2\mathbf{R}(\phi) =$$

$$= \tau^2\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} + \sigma^2\begin{bmatrix} \rho(\|x_1 - x_1\|;\phi) & \cdots & \rho(\|x_1 - x_n\|;\phi) \\ \vdots & \rho(\|x_i - x_j\|;\phi) & \vdots \\ \rho(\|x_n - x_1\|;\phi) & \cdots & \rho(\|x_n - x_n\|;\phi) \end{bmatrix}.$$

Segue que o logaritmo natural da função de verossimilhança para $(\beta;\phi)$ é dado por

$$\ln[L(\beta;\phi)] = (const.) \times (-\tfrac{1}{2})\left\{\ln|\mathbf{G}(\theta)| + (\mathbf{y} - \mathbf{F}\beta)'[\mathbf{G}(\theta)]^{-1}(\mathbf{y} - \mathbf{F}\beta)\right\}. \quad (4.11)$$

Para obtermos o máximo da expressão (4.11), primeiro eliminamos β da maximização numérica de $\ln[L(\cdot)]$. Para isto, tomamos seu estimador por máxima verossimilhança (para θ fixo)

$$\hat{\beta}(\theta) = \{\mathbf{F}'[\mathbf{G}(\theta)]^{-1}\mathbf{F}\}^{-1}\mathbf{F}'\{\mathbf{G}(\theta)\}^{-1}\mathbf{y}$$

e substituímos $\hat{\beta}(\theta)$ na expressão (4.11). Isto produzirá uma expressão reduzida do logaritmo natural da função de verossimilhança para θ,

$$\ln[L(\theta)]^* = \ln[L(\hat{\beta}(\theta),\theta)].$$

Reduções deste tipo são importantes pois permitem eliminar algebricamente parâmetros do critério a ser numericamente maximizado. Isto faz com que as rotinas de maximização sejam mais confiáveis, pois elas tendem a serem mais eficientes quando a dimensionalidade da função a ser maximizada é menor (Diggle & Ribeiro, 2000).

A Figura 4.22 mostra o variograma para os teores de areia na região estudada pelo projeto MAPEM, ajustado pelo critério de máxima verossimilhança.

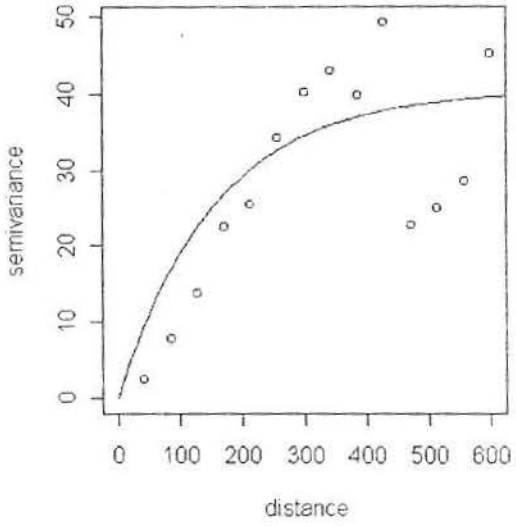

Figura 4.22 - Variograma para o teor de areia na região estudada pelo projeto MAPEM com ajuste pelo método da máxima verossimilhança de uma função de correlação exponencial com parâmetros $\hat{\theta} = (\sigma^2 = 40{,}4;\ \ \phi = 156{,}6;\ \ \tau^2 = 0)$.

Máxima Verossimilhança Restrita

Uma variação do método da máxima verossimilhança, cuja origem está no contexto de planejamento de experimentos (Patterson & Thompson[12], 1971 apud Diggle & Ribeiro, 2000), é conhecida por *método da máxima verossimilhança restrita*.

Este método consiste em transformar os dados linearmente para $\mathbf{Y}^* = \mathbf{AY}$, assumindo um modelo para $E(\mathbf{Y}) = \mathbf{F}\beta$, de tal forma que a distribuição de $\mathbf{Y}^*$ não dependa de β (Diggle & Ribeiro, 2000). A escolha da matriz **A** pode ser feita sem o conhecimento de β ou θ. Uma maneira apropriada de encontrá-la é fazer

$$\mathbf{A} = \mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'.$$

O estimador da máxima verossimilhança restrita é encontrado maximizando

$$\ln[L^*(\theta)] = (const) \times (-\tfrac{1}{2})\left\{\ln|\mathbf{G}(\theta)| + \ln|\mathbf{F}'[\mathbf{G}(\theta)]^{-1}\mathbf{F}| + (\mathbf{y} - \mathbf{F}\tilde{\beta})'[\mathbf{G}(\theta)]^{-1}(\mathbf{y} - \mathbf{F}\tilde{\beta})\right\},$$

onde $\tilde{\beta} = \hat{\beta}(\theta)$. Apesar desta expressão depender de **F** e, consequentemente, depender da correta especificação de $\mu(x)$, ela não depende da escolha de **A**.

---

[12] Patterson, H.D. e Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-54.

Segundo Diggle & Ribeiro (2000), o método da máxima verossimilhança restrita produz vícios menores para amostras pequenas e é mais sensível à especificações incorretas do modelo de $\mu(x)$ do que o método da máxima verossimilhança tradicional. A Figura 4.23 compara o ajuste do variograma para os teores de areia na região estudada pelo projeto MAPEM, realizado pelos critérios de máxima verossimilhança e máxima verossimilhança restrita.

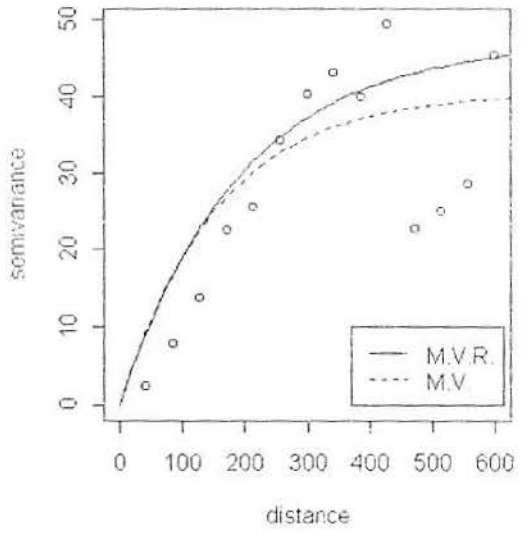

Figura 4.23 - Variograma para o teor de areia na região estudada pelo projeto MAPEM com ajuste de uma função de correlação exponencial pelo método da máxima verossimilhança restrita com parâmetros estimados $\hat{\theta} = (\sigma^2 = 47;\ \phi = 192{,}5;\ \tau^2 = 0)$ *(linha contínua) em contraste com o ajuste da mesma função pelo método da máxima verossimilhança (linha pontilhada).*

4.9. VEROSSIMILHANÇA RELATIVA (PROFILE LIKELIHOODS)

Os métodos da máxima verosimilhança baseiam-se na obtenção do máximo da superfície dada pela expressão (4.11). Entretanto, devido à dimensão desta superfície, análises de seu comportamento acabam por se tornar inviáveis.

Em virtude disto, ocorreu a proliferação de procedimentos *ad-hoc* para aproximações analíticas ou numéricas para a solução deste problema (Gamerman & Migon, 1993).

Se nossa expressão da verossimilhança é da forma $L(\sigma^2, \phi; x)$, podemos buscar a expressão para a verossimilhança marginal $L(\sigma^2; x)$ substituindo o valor de ϕ

na expressão da verossimilhança por $\hat{\phi}(\sigma^2)$, ou seja, *seu estimador da máxima verossimilhança para cada valor de* σ^2. Esta expressão marginal da verossimilhança é denominada *verossimilhança relativa* ou *profile*.

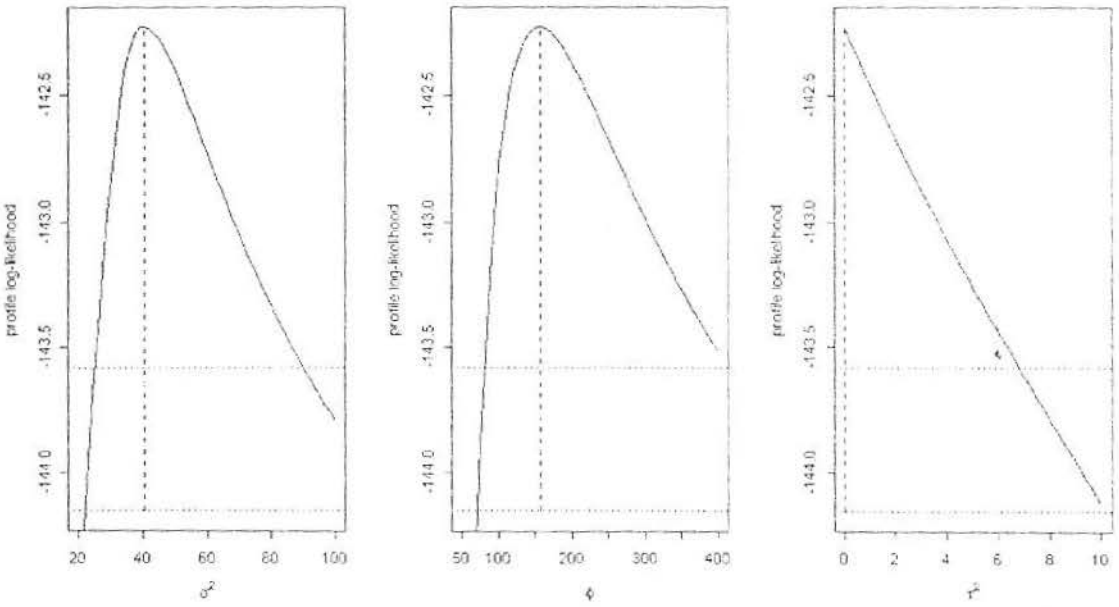

Figura 4.24 - Função de verossimilhança relativa para os parâmetros $(\sigma^2;\phi;\tau^2)$ *com os respectivos intervalos de confiança à 95% (linha pontilhada superior) e à 99% (linha pontilhada inferior) para a variável teor de areia.*

Uma observação importante a ser feita sobre a estimação de parâmetros da estrutura de covariância espacial é a de que os variogramas estimados por máxima verossimilhança e máxima verossimilhança restrita, podem ser completamente diferentes do variograma amostral. Por outro lado, os variogramas estimados por MQO e MQP tendem a ser bem próximos ao variograma amostral — o que é esperado, pois se baseiam em ajustes de curvas ao variograma amostral.

Entretanto, a proximidade entre o variograma estimado e o variograma amostral não é critério seguro para avaliar a qualidade de estimação (Diggle & Ribeiro, 2000). Podemos destacar duas grandes desvantagens dos métodos baseados em ajustes de curvas, bem como os métodos não-paramétricos baseados na suavização do variograma amostral.

A primeira grande desvantagem destes métodos de suavização é devida ao fato de que, para uma amostra de n locais, o variograma empírico terá $\frac{1}{2}n(n-1)$ valores de ordenadas v_{ij}. Isto conduzirá a uma forte *dependência* entre os v_{ij}, afetando

desfavoravelmente o desempenho dos métodos tradicionais de modelos de regressão (Diggle & Ribeiro, 2000).

Uma segunda grande desvantagem destes métodos, já descrita pela Figura 4.10 na seção 4.1.4, é o fato de que o variograma amostral pode variar drasticamente *apenas com a mudança do valor de k no cálculo do variograma amostral*.

Analisando a Figura 4.25, vemos os limites superiores e inferiores dos valores de 30 variogramas amostrais simulados, gerados a partir de processos Gaussianos com *os mesmos parâmetros estimados* pelos critérios de MQP_3 e máxima verossimilhança para o teor de areia nos dados do MAPEM. Estes processos são gerados em malhas amostrais com o mesmo número de locais amostrados existentes na área estudada pelo projeto MAPEM, que é de 47 pontos.

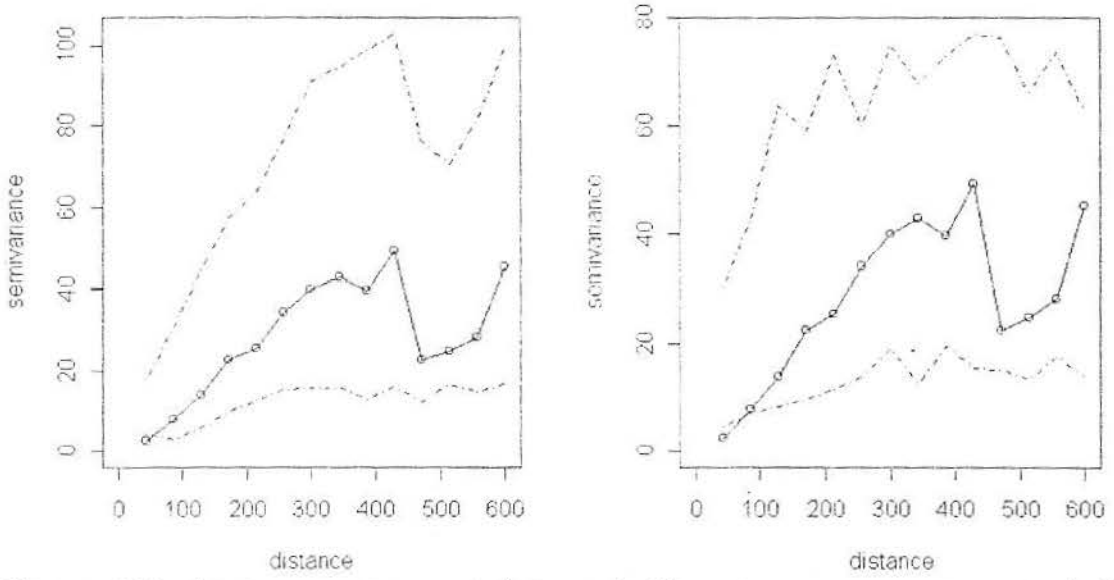

Figura 4.25 - Limites superiores e inferiores de 30 variogramas amostrais simulados em malhas de 47 pontos, cujo processo gerador dos dados é Gaussiano com parâmetros da estrutura de covariância espacial iguais aos parâmetros estimados por MQP_3 (esquerda) e estimados por máxima verossimilhança (direita).

Estes limites, ou *envelopes* (Diggle & Ribeiro, 2000), nos mostram a amplitude de variação dos variogramas amostrais gerados a partir de um mesmo processo gerador dos dados. Caso nosso variograma amostral se encontre fora destes limites, devemos realizar análises adicionais no nosso estudo, *pois se o modelo de estrutura de covariância espacial, que foi estimado, fosse realmente o modelo verdadeiro, ele dificilmente produziria um variograma amostral como o encontrado em nossos estudos.*

Um argumento normalmente aceito contra a utilização dos métodos de máxima verossimilhança em favor de métodos baseados nos critérios de mínimos quadrados é o fato de não podermos ter certeza se o nosso processo é Gaussiano. Os métodos baseados no critério de mínimos quadrados funcionam de forma razoável tanto em casos onde lidamos com processos Gaussianos quanto nos casos em que o processo não é Gaussiano (Cressie, 1991). Por outro lado, a utilização de métodos baseados na função de verossimilhança do processo gerador dos dados só apresenta bons resultados quando esta função está corretamente especificada.

4.10. COMPARAÇÕES ENTRE ESTIMAÇÕES REALIZADAS PELO CRITÉRIO DE MQP E PELO MÉTODO DA MÁXIMA VEROSSIMILHANÇA PARA PROCESSOS SIMULADOS

Apenas para efeito de comparação, vamos apresentar agora diferentes simulações de processos Gaussianos e seus respectivos variogramas estimados por MQP_2 e pela máxima verossimilhança.

De modo geral, o processo de simulação Gaussiana busca gerar valores Y_i, tal que $\mathbf{Y} \sim MVN(\mathbf{0},\Sigma)$. Um vetor $\mathbf{Y}$ com simulações para um conjunto de n pontos pode ser obtido por:

$$\mathbf{Y} = \Sigma^{1/2}\mathbf{Z},$$

onde $\mathbf{Z}$ é um vetor com n observações $Z_i \sim N(0,1)$ e $\Sigma^{1/2}$ é tal que $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})'$. Este procedimento pode ser realizado através de técnicas de decomposição da matriz Σ. Como o modelo geoestatístico Gaussiano é aditivo, se desejarmos adicionar ao modelo componentes de tendência ou efeito pepita, basta adicionar estas componentes aos valores simulados (Diggle & Ribeiro, 2000).

Os processos simulados foram gerados no pacote GeoR e os parâmetros estimados para os modelos que serão apresentados a seguir se encontram no Anexo 1.

As simulações dos modelos que serão apresentados a seguir foram realizadas de forma a dispormos de duas malhas amostrais para cada modelo, uma com 50 amostras e outra com 100 amostras. As coordenadas que referenciam estas malhas amostrais estão padronizadas e variam no intervalo [0,1]. Os modelos foram estimados pelo método da máxima verossimilhança e pelo critério de MQP_2. Como ambos os

métodos necessitam de algoritmos numéricos para realizar o procedimento de estimação, os valores dos parâmetros iniciais para interação foram os mesmos para ambos e seus valores foram exatamente os parâmetros da verdadeira estrutura de covariância espacial. Além disto, considerou-se que o modelo de família de correlação espacial era conhecido em cada caso.

i) **Modelo Esférico:** $\theta = (\sigma^2 = 0{,}1; \quad \phi = 0{,}1; \quad \tau^2 = 0)$

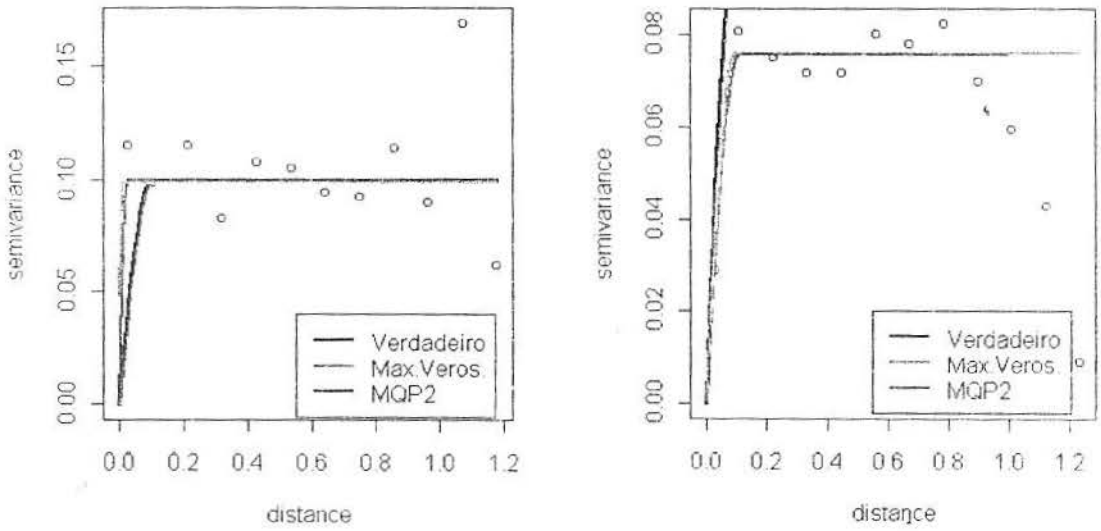

Figura 4.26 - Ajuste de variograma para malha amostral de 50 locais (à esquerda) e malha de 100 locais (à direita), realizado pelo método da máxima verossimilhança (em vermelho) e pelo critério de MQP_2 (em azul), juntamente com o verdadeiro modelo de variograma (em preto).

Nesta simulação, buscou-se analisar a estimação de parâmetros do variograma nos casos onde a variabilidade e a amplitude da dependência espacial são pequenas.

Para a amostra de 50 locais, ambos os métodos captaram apenas pequena parte da dependência espacial, nos levando a acreditar que o processo não possui esta dependência. Entretanto, com a amostra de 100 locais, a dependência espacial foi captada por ambos, sendo que apenas o patamar de variabilidade foi subestimado (ver Figura 4.26).

ii) Modelo Esférico: $\theta = (\sigma^2 = 100;\quad \phi = 0{,}4;\quad \tau^2 = 20)$

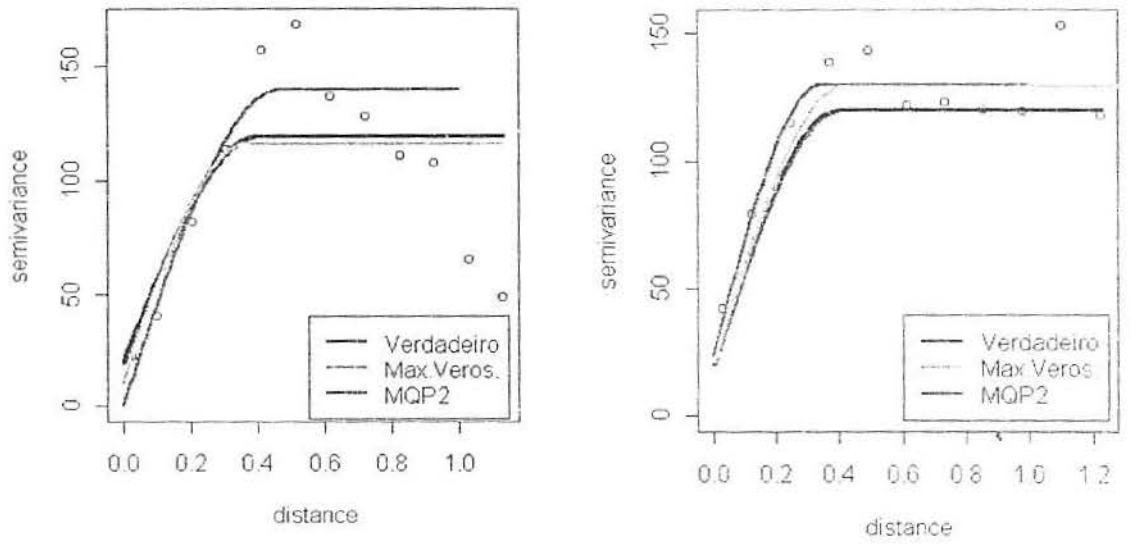

Figura 4.27 - Ajuste de variograma para malha amostral de 50 locais (à esquerda) e malha de 100 locais (à direita), realizado pelo método da máxima verossimilhança (em vermelho) e pelo critério de MQP_2 (em azul), juntamente com o verdadeiro modelo de variograma (em preto).

Analisou-se no exemplo acima outro modelo esférico com variabilidade e dependência espacial mais elevadas. Ambos os métodos obtiveram resultados satisfatórios, entretanto o variograma ajustado pelo método da máxima verossimilhança apresentou resultados muito superiores. No caso de 50 amostras, o ajuste realizado pela máxima verossimilhança foi praticamente perfeito (ver Figura 4.27).

iii)Modelo Exponencial Potência $k = 2$ (Gaussiano):

$\theta = (\sigma^2 = 50;\ \ \phi = 0{,}3;\ \ \tau^2 = 0)$

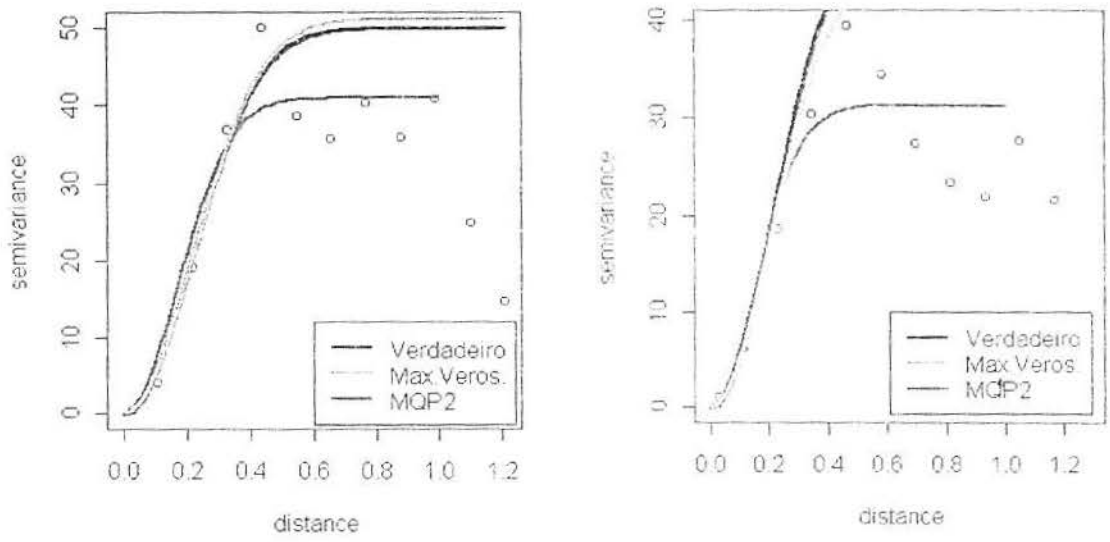

Figura 4.28 - Ajuste de variograma para malha amostral de 50 locais (à esquerda) e malha de 100 locais (à direita), realizado pelo método da máxima verossimilhança (em vermelho) e pelo critério de MQP_2 (em azul), juntamente com o verdadeiro modelo de variograma (em preto).

Neste caso, onde a função de correlação do variograma é Gaussiana, os resultados obtidos pelo método da máxima verossimilhança foram plenamente satisfatórios. É importante notar que o ajuste realizado foi extremamente preciso tanto no caso de cálculo baseado em 50 amostras, quando no caso de 100 amostras. O modelo estimado pelo critério de MQP_2 foi desfavoravelmente influenciado pelo comportamento do variograma amostral (ver Figura 4.28).

iv)Modelo Exponencial Potência com k = 1 (Exponencial):

$\theta = (\sigma^2 = 40;\quad \phi = 1;\quad \tau^2 = 5)$

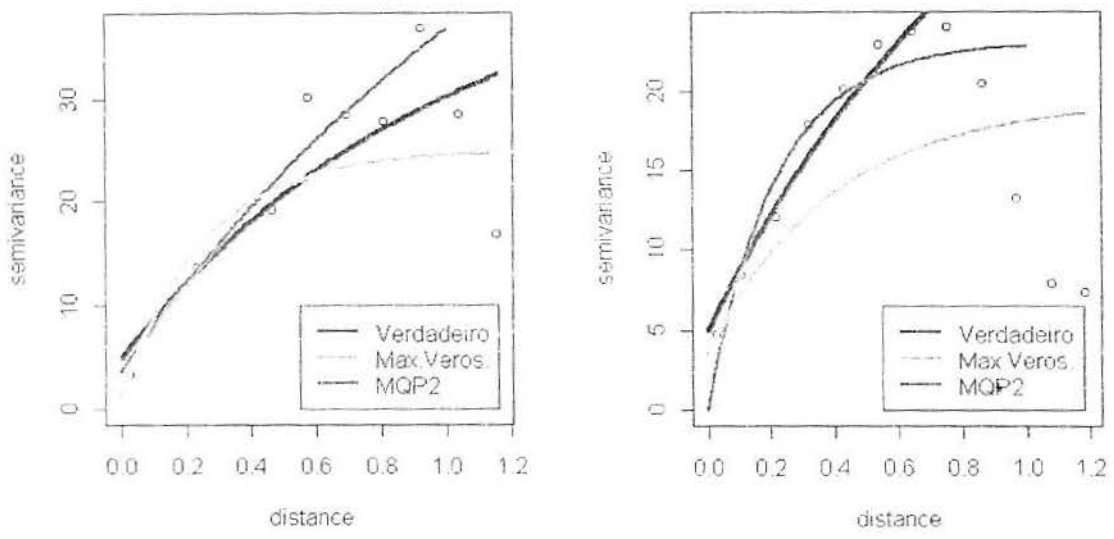

Figura 4.29 - Ajuste de variograma para malha amostral de 50 locais (à esquerda) e malha de 100 locais (à direita), realizado pelo método da máxima verossimilhança (em vermelho) e pelo critério de MQP_2 (em azul), juntamente com o verdadeiro modelo de variograma (em preto).

Nesta simulação, onde temos um modelo com alta variabilidade e amplitude de dependência espacial, tanto para 50 amostras quanto para 100 amostras, os resultados obtidos pelo critério de MQP_2 foram superiores (ver Figura 4.29).

Após esta breve análise destes modelos simulados, é importante notar que o desempenho do método da máxima verossimilhança foi geralmente superior para os casos onde a amplitude da dependência espacial é menor. A função de correlação espacial do tipo exponencial é normalmente a função que determina maior amplitude de dependência espacial. Isto acontece porque o decréscimo desta função é mais lento do que o decréscimo de outras funções derivadas da família Exponencial Potência, como a função de correlação Gaussiana (ver Figura 3.7).

Apesar dos pequenos indícios de termos uma melhor estimação, por parte dos métodos de máxima verossimilhança para os casos onde a função de correlação espacial tiver decréscimo mais rápido, *necessitaríamos de mais estudos de simulações verificando o comportamento destes métodos de estimação para diferentes situações e modelos para que possamos afirmar algo com segurança neste sentido.* É importante salientar que nestas simulações, partimos do pressuposto que:

1) A função de correlação espacial era conhecida;
2) A função de máxima verossimilhança estava corretamente especificada, pois tínhamos segurança em afirmar que os dados foram gerados por um processo Gaussiano;
3) Os valores iniciais dos parâmetros, que foram utilizados nos algoritmos numéricos para o processo de estimação, eram os mais próximos possíveis dos verdadeiros parâmetros do modelo (já que seus valores eram exatamente os parâmetros do modelo).

5. PREDIÇÃO ESPACIAL

Nos capítulos anteriores, nos preocupamos em estudar aspectos da análise Geoestatística que são importantes pré-requisitos para o principal objetivo em questão: *a predição espacial*.

Na seção 2.4 do Capítulo 2, vimos como fazer predição espacial supondo um modelo estatístico totalmente especificado. Entretanto, quando um pesquisador se depara com um certo fenômeno que pode ser analisado pelas técnicas de Geoestatística, os parâmetros do modelo a ser utilizado no processo de predição são normalmente desconhecidos.

A partir disto, o pesquisador vê-se obrigado a estimar estes parâmetros desconhecidos — normalmente parâmetros que especificam a estrutura de covariância espacial — e considerá-los como os verdadeiros parâmetros populacionais a fim de realizar o processo de predição. Infelizmente, este tipo de prática adiciona erros nas predições e nas respectivas variâncias de predições.

Neste Capítulo, vamos apresentar o método geral de predição em Geoestatística, realizando a predição espacial para dados obtidos do projeto MAPEM, além de dados obtidos por processo de simulação, utilizando parâmetros estimados por critérios baseados em mínimos quadrados e máxima verossimilhança.

5.1. O PREDITOR DE $S(x)$

No Capítulo 2, vimos que a predição de $T = S(x)$ para qualquer $x \in D$, realizada através de um preditor $\hat{T} = f(\mathbf{Y})$, possui erro quadrático médio mínimo quando

$$\hat{T} = E(T / \mathbf{Y}).$$

Logo, o preditor $\hat{T} = E(T / \mathbf{Y})$ é *preditor ótimo* de T, segundo o critério de menor erro quadrático médio. Entretanto, o preditor $\hat{T} = E(T / \mathbf{Y})$ *nem sempre é linear em* $\mathbf{Y}$.

O uso de preditores lineares é justificado quando possuímos informações incompletas sobre a distribuição do fenômeno em estudo (Gamerman & Migon, 1993).

No contexto de Geoestatística, o cálculo da média e da variância do erro de predição só são possíveis se conhecemos a distribuição do processo $S(x)$. Entretanto, como possuímos normalmente apenas uma única realização do processo $S(x)$, não podemos inferir sua distribuição precisamente. Por este motivo, nos restringimos à escolha de preditores pertencentes à classe dos estimadores lineares, que sempre nos possibilitarão calcular a média e a variância do erro de predição a partir do variograma ou do covariograma (Journel & Huijbregts, 1978).

A partir disto, buscamos o *preditor linear ótimo em* $\mathbf{Y}$, a fim de predizer os valores de $T = S(x)$. Este preditor, é da forma

$$\hat{T} = \hat{S}(x) = \sum_{i=1}^{n} w_i(x) Y_i .$$

onde os pesos $w_i(x)$ são derivados da média estimada e da estrutura de covariância dos dados.

A grande vantagem de trabalharmos supondo um processo Gaussiano $S(x)$ está no fato de que o preditor ótimo de T baseado em $\mathbf{Y}$, ou seja $\hat{T} = E(T / \mathbf{Y})$, é também o preditor *linear* ótimo de T. Considerando que $E[S(x)] = \mu \ (\forall x \in D)$, vimos no Capítulo 2 que este preditor é dado por:

$$\hat{T} = \mu + \sigma^2 \mathbf{r}'(\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1}(\mathbf{Y} - \mu \mathbf{1})$$

e possui variância de predição

$$Var(T / \mathbf{Y}) = \sigma^2 - \sigma^2 \mathbf{r}'(\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1} \sigma^2 \mathbf{r} .$$

Além deste fato, assumir que nosso processo é Gaussiano traz ainda outras grandes vantagens (Cressie, 1991):

i) Homoscedasticidade condicional: A variância $Var(T / \mathbf{Y})$ não depende de $\mathbf{Y}$, ou seja, só depende da configuração espacial dos locais x na amostra e da estrutura de covariância espacial;

ii) O erro quadrático médio de $\hat{T} = E(T / \mathbf{Y})$ é igual à variância de predição $Var(T / \mathbf{Y})$.

O método de predição espacial que utiliza o preditor $\hat{T} = E(T / \mathbf{Y})$ a fim de construir uma superfície $\hat{S}(x)$ é chamado de *krigeagem*.

5.2. KRIGEAGEM

Krigeagem é o método de predição baseado na minimização do erro quadrático médio que normalmente depende das propriedades de 2ª ordem do processo $S(x)$.

Normalmente, nosso objetivo é predizer o valor de $S(x)$ ou seu valor médio em uma determinada região, ou seja,

$$T = \frac{1}{|B|}\int_B S(x)dx,$$

onde $|B|$ é a área da região B $(B \in D)$. Pela linearidade do operador esperança, temos que o preditor de T é dado por (Diggle & Ribeiro, 2000):

$$\hat{T} = \frac{1}{|B|}\int_B \hat{S}(x)dx.$$

Suponha que os dados $\mathbf{Y} = (Y_1, \ldots, Y_n)$ foram gerados por um modelo Gaussiano linear,

$$Y_i = \mu(x_i) + S(x_i) + Z_i : i = 1, \ldots, n,$$

onde $S(x_i)$ é um processo Gaussiano estacionário com média zero, variância σ^2, função de correlação $\rho(h;\phi)$, $\mu(x_i)$ determinado por um modelo de regressão baseado em k funções de variáveis referenciadas observadas

$$\mu(x_i) = \sum_{k=1}^{p} f_k(x_i)\beta_k$$

e $Z_i \sim N(0, \tau^2)$ é um ruído branco. Então, segue que $\mathbf{Y}$ tem distribuição Normal Multivariada

$$\mathbf{Y} \sim MVN(\mathbf{F}\beta, \mathbf{G}(\theta)),$$

onde

- $\theta = (\tau^2, \sigma^2, \phi)$
- $E(\mathbf{Y}) = \mathbf{F}\beta = [\text{matriz de delineamento } n \times p]\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$
- $Var(\mathbf{Y}) = \mathbf{G}(\theta) = \tau^2\mathbf{I} + \sigma^2\mathbf{R}(\phi) =$

$$= \tau^2\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} + \sigma^2\begin{bmatrix} \rho(\|x_1 - x_1\|;\phi) & \cdots & \rho(\|x_1 - x_n\|;\phi) \\ \vdots & \rho(\|x_i - x_j\|;\phi) & \vdots \\ \rho(\|x_n - x_1\|;\phi) & \cdots & \rho(\|x_n - x_n\|;\phi) \end{bmatrix}$$

onde $\rho(\|x_i - x_j\|) = Corr\{S(x_i), S(x_j)\} = Corr\{S(x), S(x+h)\}$

- $Y_i : i = 1, \ldots, n$ são mutuamente independentes, condicionais à $S(x)$.

A partir deste modelo, com base na seção 2.4, temos que o preditor de $S(x)$ é

$$\hat{T} = \mu(x) + \sigma^2 \mathbf{r}'(\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1}(\mathbf{Y} - \mathbf{F}\beta) \tag{5.1}$$

com variância de predição

$$Var(T / \mathbf{Y}) = \sigma^2 - \sigma^2 \mathbf{r}'(\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1} \sigma^2 \mathbf{r} \tag{5.2}$$

onde $\mathbf{r} = \begin{bmatrix} \rho(\|x - x_1\|) \\ \vdots \\ \rho(\|x - x_n\|) \end{bmatrix}$.

O processo de realizar a predição através da expressão dada por (5.1), construindo um mapa de superfície para $\hat{S}(x)$, é conhecido como *krigeagem simples*.

A partir da expressão (5.2), podemos construir um mapa da precisão da predição, tomando o desvio-padrão

$$DP(x) = \sqrt{Var[\hat{S}(x) / \mathbf{Y}]}. \tag{5.3}$$

Embora o preditor da krigeagem possa ser derivado como o preditor linear ótimo sem realizar referência explícita à um modelo Gaussiano, a predição espacial baseada em $E(T / \mathbf{Y})$ é linear *somente* sob suposição Gaussiana (Diggle, Moyeed e Twan, 1998).

Na krigeagem simples, normalmente substituímos os parâmetros que definem a estrutura de covariância espacial por suas estimativas e assumimos que a média é conhecida a priori. Quando não desejamos — ou não podemos — supor que a média é conhecida, nos voltamos aos procedimento de *krigeagem ordinária* e *krigeagem universal*, que podem ser considerados como casos particulares da predição bayesiana (Diggle & Ribeiro, 2000). Nesta monografia, será apresentada apenas a predição espacial baseada nos métodos de krigeagem simples. Para referências a outros métodos de krigeagem, sob a suposição de modelos estatísticos explícitos, recomenda-se Diggle & Ribeiro (2000). Os métodos de predição baseados em krigeagem simples, ordinária e universal são classificados como métodos de *krigeagem linear*. Para referências à métodos de krigeagem não-linear, ver Diggle, Moyeed e Twan (1998).

Exemplo de Krigeagem Simples

Suponha que desejamos estimar o teor de uma certa substância existente em um determinado local x_0 com base em amostras do teor desta substância, observadas em três locais vizinhos x_1, x_2 e x_3.

A Figura 5.1 ilustra a localização das amostras e o local onde desejamos predizer o teor da substância. As distâncias entre os locais e os seus respectivos valores são dados por:

Tabela 5.1 - Distâncias entre os locais x.

Local A	Local B	Distância
x_0	x_1	15,00 metros
x_0	x_2	24,00 metros
x_0	x_3	10,00 metros
x_1	x_2	39,00 metros
x_1	x_3	18,02 metros
x_2	x_3	26,00 metros

Tabela 5.2 - Valores observados.

Local	Valores (Y)
x_1	64
x_2	70
x_3	47

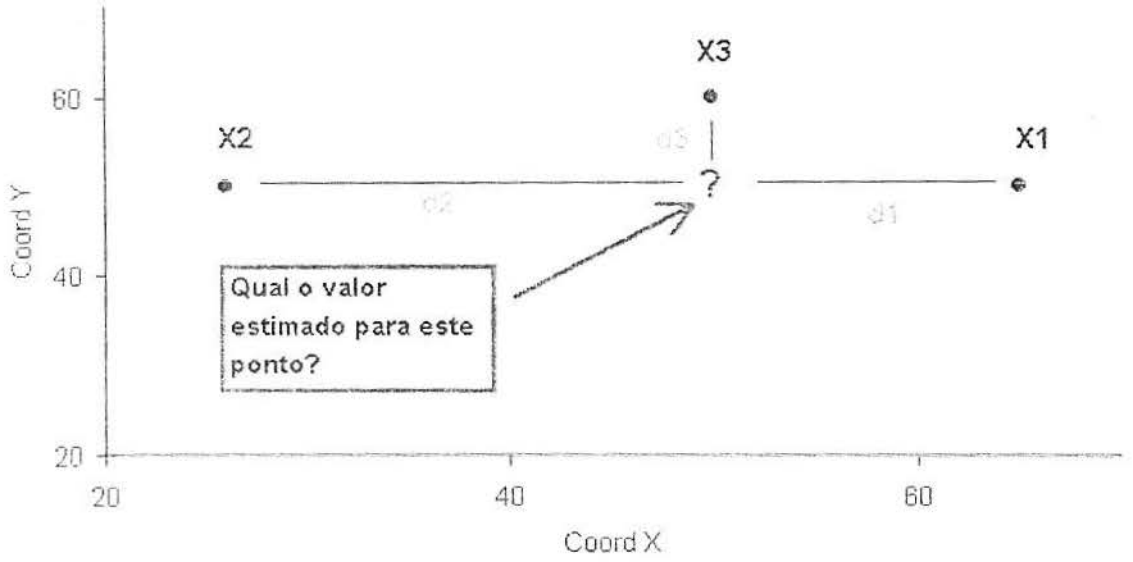

Figura 5.1 - Localização das amostras x_1, x_2 e x_3 e do local onde desejamos estimar o teor da substância.

Para realizar a predição, supomos que os parâmetros que definem a estrutura de covariância espacial são dados por

$\mu = 60$

$\tau^2 = 10$,

$\sigma^2 = 110$

com função de correlação espacial dada por

$$\rho(h;\phi) = Esférica(h;20) = \begin{cases} 1 - \frac{3}{2}(h/20) + \frac{1}{2}(h/20)^3 : 0 \le h \le 20 \\ 0 \qquad\qquad\qquad\qquad\quad : h \ge 20 \end{cases}$$

A Tabela 5.3 fornece os valores desta função de correlação espacial para todas as distâncias consideradas neste exemplo.

Tabela 5.1 - Distâncias entre os locais x.

Local A	Local B	Distância	$\rho(h)$
x_0	x_1	15,00 metros	0,0859
x_0	x_2	24,00 metros	0,0000
x_0	x_3	10,00 metros	0,3125
x_1	x_2	39,00 metros	0,0000
x_1	x_3	18,02 metros	0,0140
x_2	x_3	26,00 metros	0,0000

A partir disto, podemos encontrar o valor de $\hat{S}(x_0)$ através da expressão (5.2):

$$\hat{S}(x_0) = \mu(x) + \sigma^2 \mathbf{r}'(\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1}(\mathbf{Y} - \mathbf{F}\beta).$$

Entretanto, como $\mu(x) = \mu$, podemos escrever (5.2) de maneira simplificada como

$$\hat{S}(x_0) = \mu + \sigma^2 \mathbf{r}'(\tau^2 \mathbf{I} + \sigma^2 \mathbf{R})^{-1}(\mathbf{Y} - \mu\mathbf{1}). \qquad (5.4)$$

Procedendo aos cálculos, temos

$$\hat{S}(x_0) = \mu + \sigma^2 \left[\rho(\|x - x_1\|) \quad \rho(\|x - x_2\|) \quad \rho(\|x - x_3\|)\right] \times$$

$$\times \left\{ \left(\tau^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} \rho(\|x_1 - x_1\|) & \rho(\|x_1 - x_2\|) & \rho(\|x_1 - x_3\|) \\ \rho(\|x_2 - x_1\|) & \rho(\|x_2 - x_2\|) & \rho(\|x_2 - x_3\|) \\ \rho(\|x_3 - x_1\|) & \rho(\|x_3 - x_2\|) & \rho(\|x_3 - x_3\|) \end{bmatrix} \right) \right\}^{-1} \times$$

$$\times \left(\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} - \mu \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)$$

$$\hat{S}(x_0) = 60 + 110\left[0{,}0859 \quad 0 \quad 0{,}3125\right] \times$$

$$\times \left\{ \left(10 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + 110 \begin{bmatrix} 1 & 0 & 0{,}014 \\ 0 & 1 & 0 \\ 0{,}014 & 0 & 1 \end{bmatrix} \right) \right\}^{-1} \left(\begin{bmatrix} 64 \\ 70 \\ 47 \end{bmatrix} - 60 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)$$

$$\hat{S}(x_0) = 60 + \left[-3{,}29286\right] = \mathbf{56{,}70714}$$

Logo, a nossa estimativa do teor da substância para o local x_0, com base em três amostras, é igual à 56,70714.

Se desejamos obter um intervalo de confiança, devemos calcular o valor da variância de predição dado por (5.2)

$$Var[\hat{S}(x_0) / \mathbf{Y}] = \sigma^2 - \sigma^2 \left[\rho(\|x - x_1\|) \quad \rho(\|x - x_2\|) \quad \rho(\|x - x_3\|)\right] \times$$

$$\times \left\{ \left(\tau^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} \rho(\|x_1 - x_1\|) & \rho(\|x_1 - x_2\|) & \rho(\|x_1 - x_3\|) \\ \rho(\|x_2 - x_1\|) & \rho(\|x_2 - x_2\|) & \rho(\|x_2 - x_3\|) \\ \rho(\|x_3 - x_1\|) & \rho(\|x_3 - x_2\|) & \rho(\|x_3 - x_3\|) \end{bmatrix} \right) \right\}^{-1} \times$$

$$\times \sigma^2 \begin{bmatrix} \rho(\|x - x_1\|) \\ \rho(\|x - x_2\|) \\ \rho(\|x - x_3\|) \end{bmatrix}$$

$$Var[\hat{S}(x_0)/\mathbf{Y}] = 110 - 110[0{,}0859 \quad 0 \quad 0{,}3125]\times$$

$$\times\left\{\left(10\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}+110\begin{bmatrix}1&0&0{,}014\\0&1&0\\0{,}014&0&1\end{bmatrix}\right)\right\}^{-1}\times 110\begin{bmatrix}0{,}0859\\0\\0{,}3125\end{bmatrix}$$

$$Var[\hat{S}(x_0)/\mathbf{Y}] = 110 - [10{,}52345] = \mathbf{99{,}47655}$$

Logo, o desvio padrão da predição é dado por

$$\sqrt{Var[\hat{S}(x_0)/\mathbf{Y}]} = \sqrt{99{,}47655} \cong \mathbf{9{,}974}$$

Com base neste resultado, obtemos um intervalo com 95% de confiança para o teor da substância no local x_0:

$$IC\ 95\% \rightarrow (56{,}707 - 1{,}96\times 9{,}974\ ;\ 56{,}707 + 1{,}96\times 9{,}974)$$
$$\rightarrow \mathbf{(37{,}15\ ;\ 76{,}25)}$$

5.3. APLICAÇÃO DAS TÉCNICAS DE KRIGEAGEM NA ANÁLISE DE DADOS DO PROJETO MAPEM

Ao longo desta monografia utilizamos dados fornecidos pelo projeto MAPEM (ver Anexo 2) a fim de estudar as características da estrutura de covariância espacial. Como prosseguimento, vamos realizar predições para locais não amostrados utilizando os resultados obtidos no capítulo anterior. A variável que será analisada é o *teor de areia* dos sedimentos. As amostras foram obtidas através de uma análise de granulometria nos 47 locais amostrados.

Como apresentado no capítulo anterior, o variograma amostral do teor de areia tem o seguinte formato:

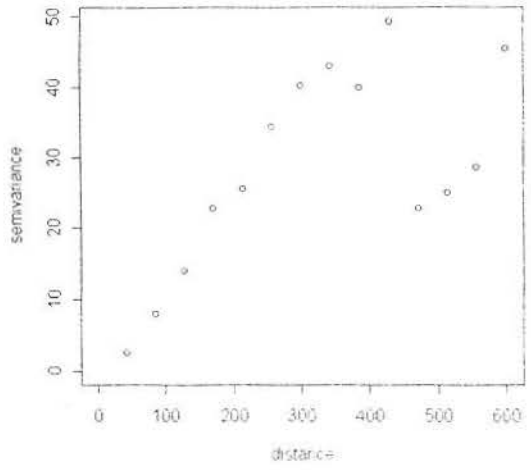

Figura 5.2 - Variograma amostral para o teor de areia.

Após uma análise do comportamento do variograma para diferentes distâncias, como mostra a Figura 5.3, concluímos que é razoável supor isotropia nos dados.

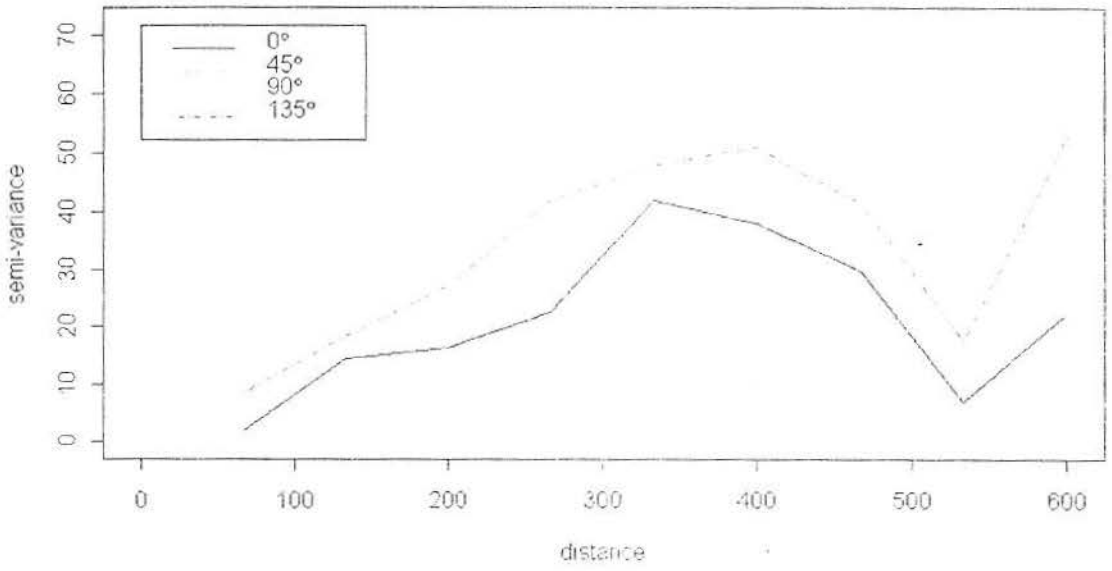

Figura 5.3 - Variograma amostral calculado para as direções 0°, 45°, 90° e 135°.

Quanto ao modelo de estrutura de covariância espacial, utilizamos a função de correlação exponencial e estimamos os parâmetros através de diversos critérios. Vamos agora analisar os resultados obtidos pelos critérios de MQO, MQP_2 e por máxima verossimilhança. Os variogramas ajustados se encontram na Figura 5.4:

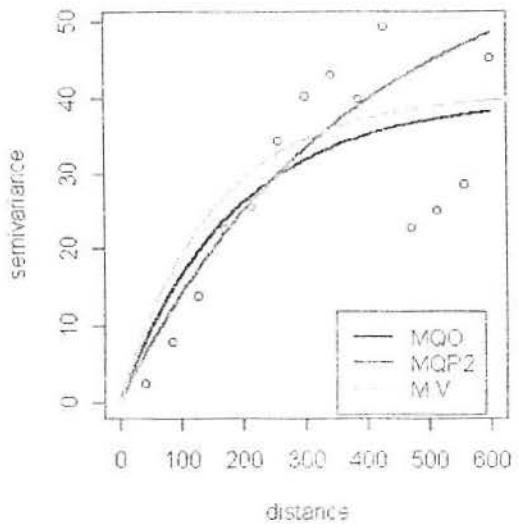

Figura 5.4 - Variogramas ajustados para a variável teor de areia segundo os critérios de MQO, MQP_2 e máxima verossimilhança.

Os parâmetros estimados por estes três critérios são dados a seguir:

Mínimos Quadrados Ordinários: $\hat{\theta} = (\sigma^2 = 39{,}7; \quad \phi = 187{,}1; \quad \tau^2 = 0)$

Mínimos Quadrados Ponderados (2): $\hat{\theta} = (\sigma^2 = 41{,}2; \quad \phi = 186{,}5; \quad \tau^2 = 0)$

Máxima Verossimilhança: $\hat{\theta} = (\sigma^2 = 40{,}4; \quad \phi = 156{,}6; \quad \tau^2 = 0)$

A fim de realizar a predição espacial por krigeagem simples, devemos supor que a média do teor de areia em toda a região de estudo é conhecida.

Quando utilizamos o critério da máxima verossimilhança, obtemos estimativas para os parâmetros da estrutura de covariância espacial juntamente com o parâmetro (ou parâmetros, quando temos média não constante na região) que definem a média do processo. Assim, podemos utilizar estes resultados como estimativas da verdadeira média. Entretanto, quando utilizamos algum critério baseado em mínimos quadrados, não temos esta estimativa.

Para evitar este tipo de problema, muitos textos recomendam o uso de técnicas de predição que não exijam o conhecimento direto dos parâmetros que definem a média, como krigeagem ordinária e krigeagem universal. Mas, como estamos interessados na aplicação das técnicas de krigeagem simples, vamos supor que a média amostral dos dados é uma razoável estimativa para a média do processo subjacente. Esta suposição encontra fundamento na medida que supomos que nosso processo é estacionário e utilizando os conceitos de ergodicidade descritos no Capítulo 2. Assim, as médias

utilizadas neste processo de predição serão iguais a 11,62 (para MQO e MQP_2) e 10,62 (para máxima verossimilhança).

A fim de obtermos uma superfície praticamente contínua sobre a região de estudo, realizou-se a predição com base nas 47 amostras do teor de areia para 10.000 locais x_i.

Em virtude dos modelos ajustados por MQO e MQP_2 serem muito semelhantes, diferenças visuais nos resultados preditos são praticamente imperceptíveis (Figura 5.5). Por este motivo e também pelo fato de que o método de MQP_2 é um dos critérios de ajuste de modelos mais utilizados dentro da Geoestatística, vamos comparar os resultados apenas deste critério em relação aos resultados obtidos pelo modelo ajustado por máxima verossimilhança.

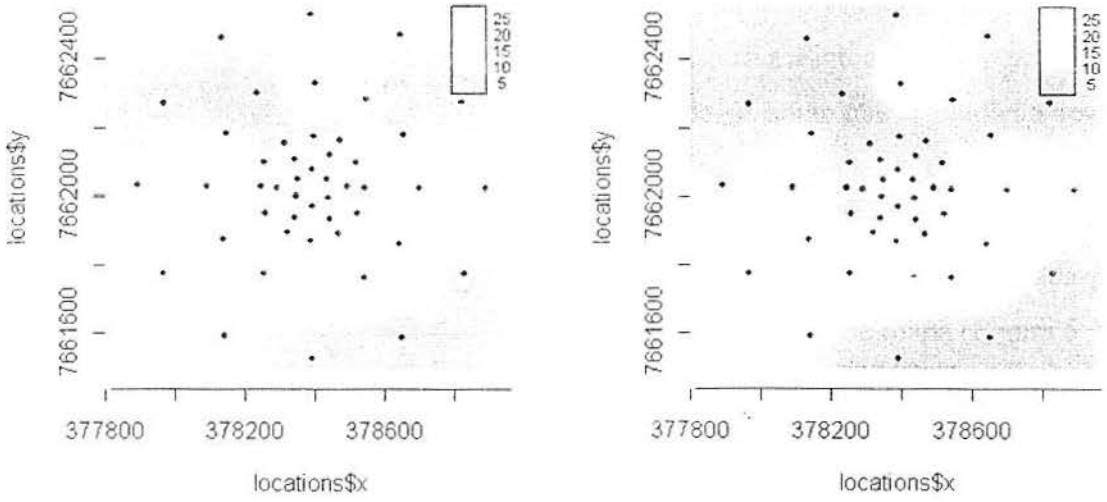

Figura 5.5 - Predição espacial por krigeagem simples para o teor de areia com modelos ajustados por MQO (esquerda) e MQP_2 (direita)

Para realizarmos uma análise mais aprofundada nos resultados, podemos diminuir o número de cores da imagem. Apesar disto, devemos salientar que o número elevado de cores é importante em uma análise visual dos resultados, pois nos fornece uma impressão maior de continuidade no mapa. Diminuindo o número de cores, podemos indicar algumas diferenças entre os resultados obtidos pelos modelos ajustados por MQP_2 e máxima verossimilhança (Figura 5.6).

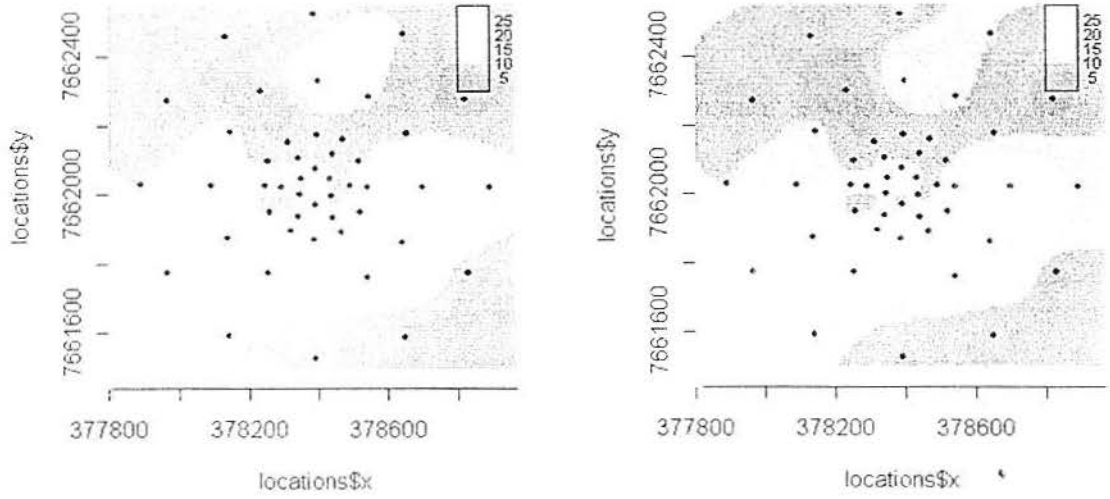

Figura 5.6. Predição espacial por krigeagem simples para o teor de areia com modelos ajustados por MQP_2 (esquerda) e máxima verossimilhança (direita).

Analisando a Figura 5.6, podemos perceber diferenças significativas entre os resultados. Pode-se notar que a predição realizada utilizando o modelo ajustado por MQP_2 possui uma amplitude de dependência entre os dados maior do que na predição realizada pelo modelo ajustado por máxima verossimilhança. Este fato se torna evidente analisando as "manchas" de cores no mapa predito pelo modelo ajustado por MQP_2, principalmente nas bordas do mapa (Figura 5.7).

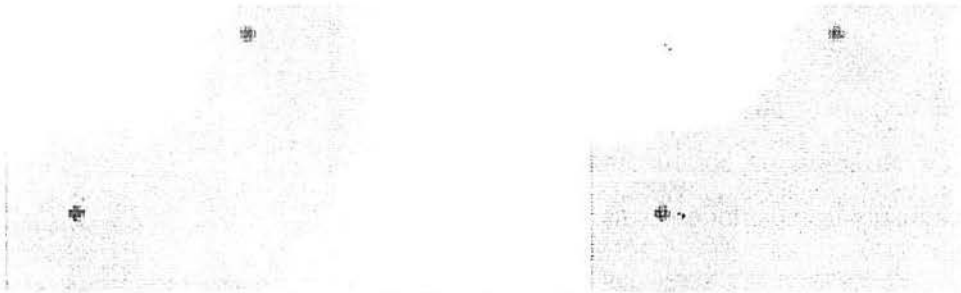

Figura 5.7 - Comparação dos resultados de predição espacial por krigeagem simples para o teor de areia no sudeste da região com modelos ajustados por MQP_2 (esquerda) e máxima verossimilhança (direita).

Analisando a Figura 5.7, podemos verificar que existe uma *zona de influência conjunta* dos dois valores Y_i nos pontos localizados entre eles apenas no caso da predição realizada pelo modelo ajustado por MQP_2.

Apesar de podermos detectar diferenças entre as predições realizadas pelos dois modelos, este não é o foco da análise. Nosso objetivo agora é saber qual dos dois modelos realizou a predição mais próxima da realidade.

Infelizmente, analisando a predição que estes modelos realizaram para os locais onde possuímos valores observados não foi satisfatória. Isto pode ser comprovado analisando o gráfico de dispersão entre os valores preditos e os valores observados (Figura 5.8).

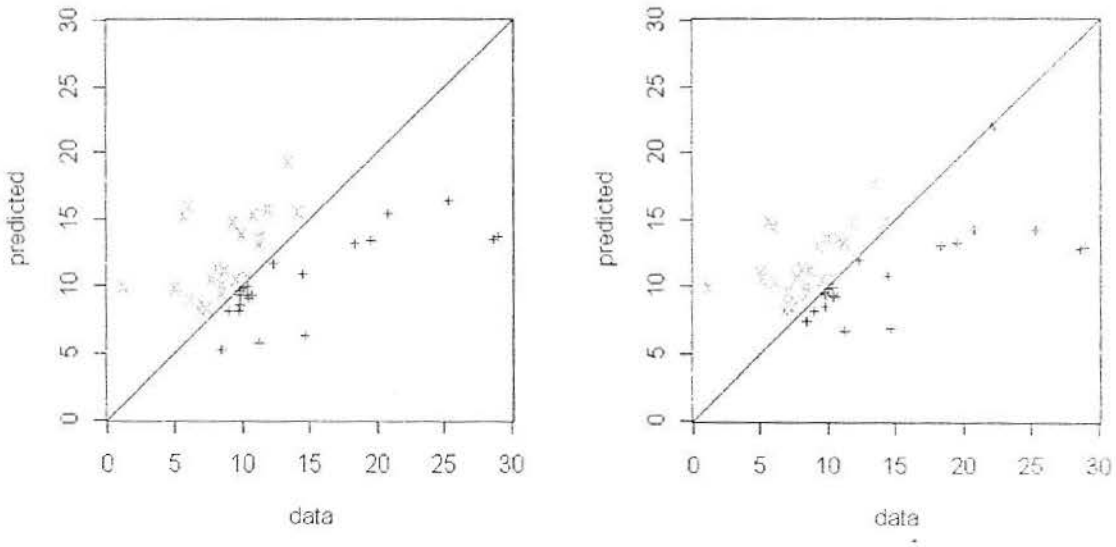

Figura 5.8 - Diagrama de dispersão entre valores observados do teor de areia e valores preditos pelo modelo ajustado por MQP_2 (esquerda) e por máxima verossimilhança (direita).

Outra maneira de visualizar esta informação é adicionar os valores originais dos locais amostrados ao mapa de valores preditos (Figura 5.9). Uma análise deste gráfico reforça as conclusões anteriores.

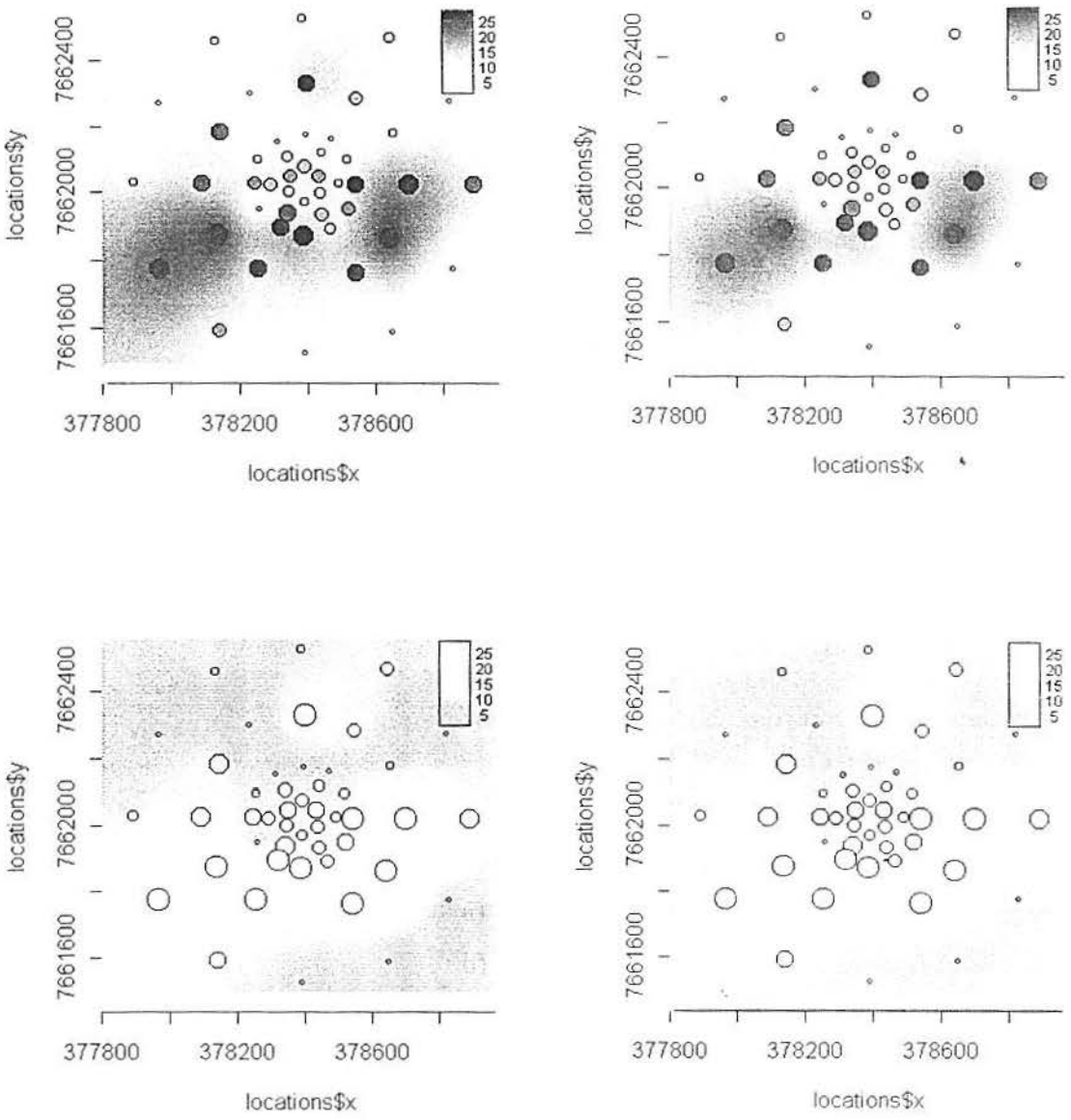

Figura 5.9 - Comparação entre valores observados do teor de areia (círculos) e valores preditos pelo modelo ajustado por MQP_2 (esquerda) e por máxima verossimilhança (direita) para duas escalas de cores diferentes.

Apenas para complementar a análise, vamos supor que o modelo ajustado por máxima verossimilhança é o verdadeiro modelo de estrutura de covariância espacial do teor de areia e, a partir disto, vamos calcular os valores da variância de predição. O mapa com as variâncias de predição é ilustrado na Figura 5.10.

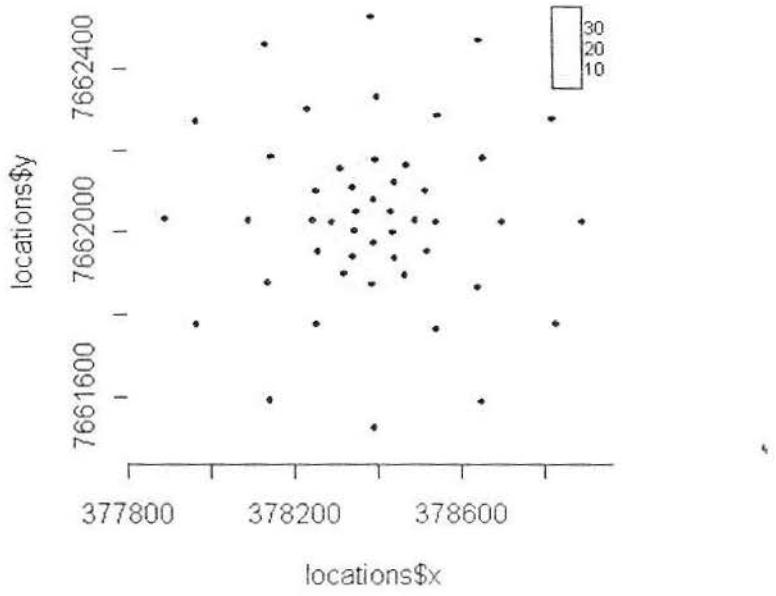

Figura 5.10 - Variância de predição para o teor de areia supondo que o modelo da estrutura de covariância espacial estimado pelo método da máxima verossimilhança é o verdadeiro modelo de estrutura de covariância espacial.

É importante salientar que esta discrepância entre os dados originais e os valores preditos é normalmente causada por dois tipos de problemas:

a) Média não constante na região de estudo;

b) O processo gerador $S(x)$ não ser Gaussiano.

Outro possível e comum fator causador deste tipo de discrepância é o efeito causado por valores atípicos. Entretanto, após algumas análises nos dados deste exemplo, não parece razoável supor que o valor do teor de areia de algum dos 47 pontos possa ser considerado como valor atípico.

Como não sabemos ao certo se os nossos dados provém de um processo Gaussiano multivariado, julgamos razoável tentar uma transformação nos dados. A transformação utilizada será a transformação de Box & Cox. Segundo Diggle & Ribeiro (2000), a transformação de Box & Cox é recomendada para os casos onde possuímos variáveis com distribuições assimétricas. Além disso, esta transformação é útil para contornar uma possível heteroscedasticidade no processo (Bailey & Gatrell, 2000). A partir desta transformação, o modelo é reformulado para (Diggle & Ribeiro, 2000):

i) Uma variável $Y^* \sim MVN(F\beta, G(\theta))$ descrita conforme (4.9);

ii) Os dados $y=(y_1,\ldots,y_n)$ são gerados por uma transformação do modelo Gaussiano $Y=h_\lambda^{-1}(Y^*)$, onde:

$$Y_i^* = h_\lambda(Y) = \begin{cases} \dfrac{(y_i)^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log(y_i) & \text{se } \lambda = 0 \end{cases}$$

A partir deste modelo, podemos estimar o valor de λ através da função de máxima verossimilhança baseada nos dados originais $y=(y_1,\ldots,y_n)$:

$$\ln(L(\beta,\theta,\lambda) = -\tfrac{1}{2}\{\ln|G(\theta)| + [h_\lambda(y) - F\beta]'[G(\theta)]^{-1}[h_\lambda(y) - F\beta]\} + \sum_{i=1}^{n} \ln\left((y_i)^{\lambda-1}\right).$$

Analisando a verossimilhança relativa (Figura 5.11), vemos que o valor que maximiza a função de verossimilhança é $\lambda = 0{,}3655$.

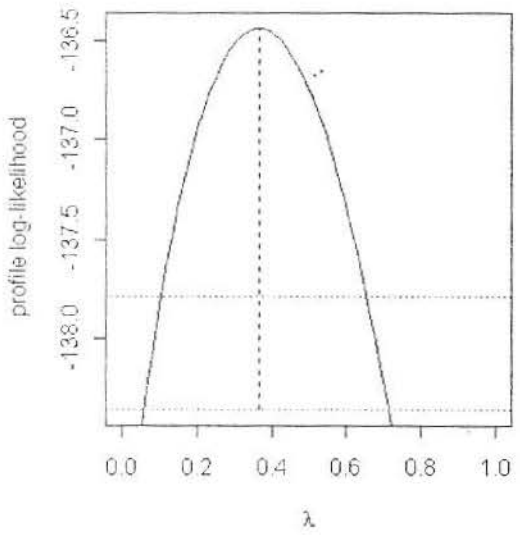

Figura 5.11 - Verossimilhança relativa para o valor do parâmetro de transformação λ.

Entretanto, a melhora na predição não justifica a transformação. A figura 5.12 confirma que as predições são semelhantes às realizadas pelos modelos ajustados aos dados originais. Neste caso, o modelo foi ajustado aos dados transformados por máxima verossimilhança. A função de correlação ajustada é do tipo Gaussiana, com parâmetros estimados $\hat{\theta} = (\sigma^2 = 1{,}652;\quad \phi = 144{,}39;\quad \tau^2 = 0;\quad \lambda = 0{,}3365)$.

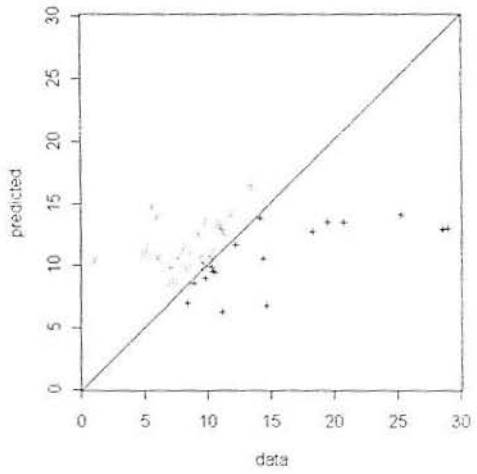

Figura 5.12 - Diagrama de dispersão entre valores observados do teor de areia e valores preditos pelo modelo ajustado por máxima verossimilhança considerando os dados transformados.

Uma hipótese plausível para estas predições "pobres", do ponto de vista estatístico, é que a média deve ser não-constante para pequenas regiões dentro da nossa região de estudo. Entretanto, este tipo de efeito não pode ser captado através de uma análise de superfícies com funções lineares, tampouco com funções quadráticas.

5.4. APLICAÇÃO DAS TÉCNICAS DE KRIGEAGEM NA ANÁLISE DE UM PROCESSO GAUSSIANO SIMULADO

No Capítulo 4 foram realizadas simulações a fim de analisar o processo de estimação dos parâmetros da estrutura de covariância espacial realizado por MQP_2 e máxima verossimilhança. A fim de que possamos avaliar o desempenho de um processo de predição por krigeagem simples que forneça resultados ótimos, sob o ponto de vista de maior precisão, vamos utilizar o processo Gaussiano simulado para 50 amostras dado pelo modelo ii) da seção 4.10.

Analisando a Figura 4.27 do capítulo anterior, podemos perceber que o modelo ajustado pelo critério da máxima verossimilhança está muito próximo do verdadeiro modelo que gerou os dados. Isto significa, que uma predição por krigeagem simples a partir deste modelo geraria resultados com a maior precisão possível.

A partir deste modelo, realizamos o processo de predição obtendo os resultados dados pela Figura 5.13, cujas variâncias de predição são dadas pela Figura 5.14.

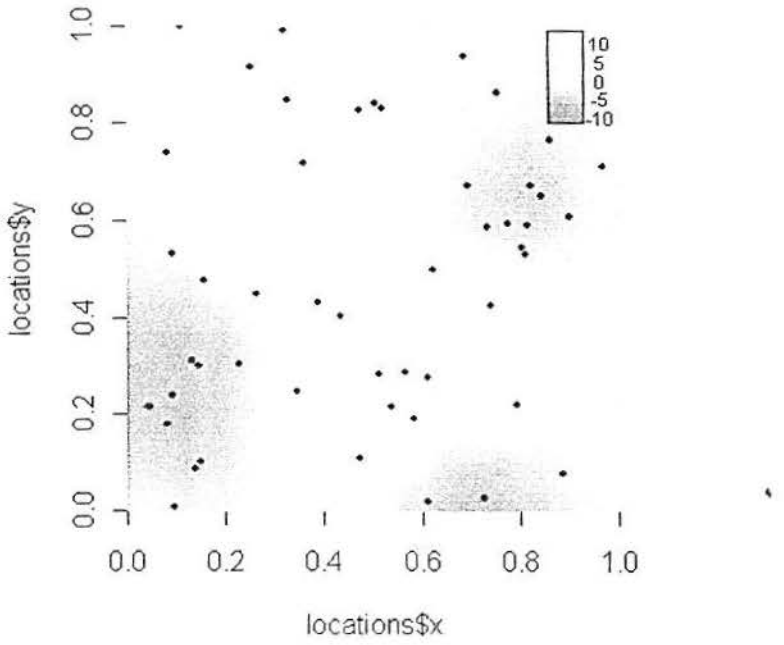

Figura 5.13 - Predição espacial por krigeagem simples para processo simulado utilizando modelo ajustado por máxima verossimilhança.

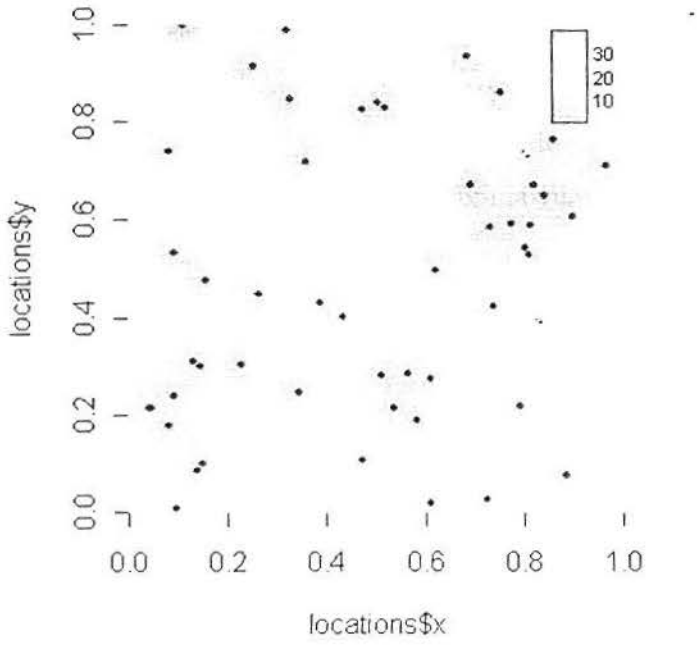

Figura 5.14 - Variâncias de predição para processo simulado utilizando modelo ajustado por máxima verossimilhança.

Para ilustrar a qualidade do ajuste, adicionaremos os valores observados ao mapa dos valores preditos (Figura 5.15). A partir de uma análise visual é possível verificar que o modelo ajustado realizou uma boa predição para os locais previamente observados.

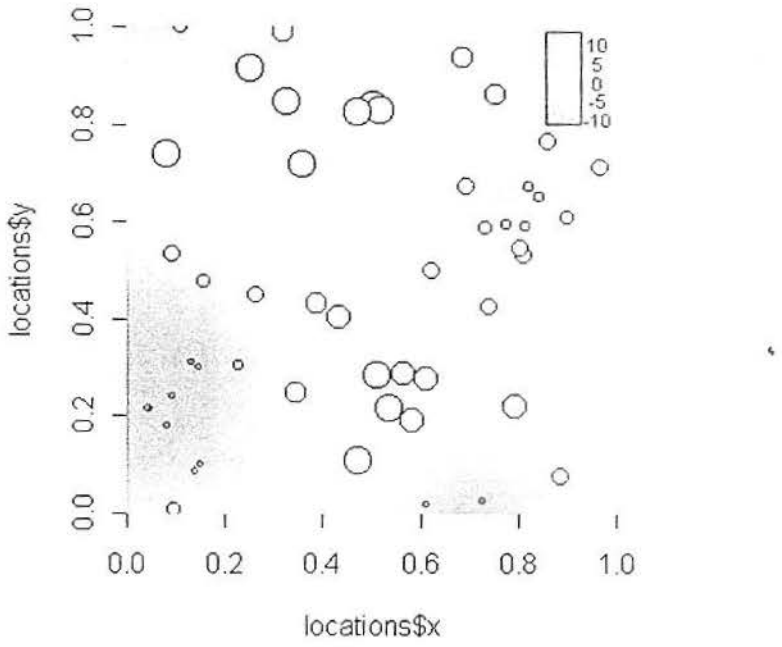

Figura 5.15 - Comparação entre valores observados do processo simulado (círculos) e valores preditos pelo modelo ajustado por máxima verossimilhança.

Para fins de comparação, a Figura 5.16 apresenta o gráfico de dispersão entre os valores originais do processo simulado e os valores preditos, utilizando o modelo ajustado por máxima verossimilhança e o modelo ajustado por MQP_2, cujos parâmetros são apresentados na seção 4.10 do capítulo anterior.

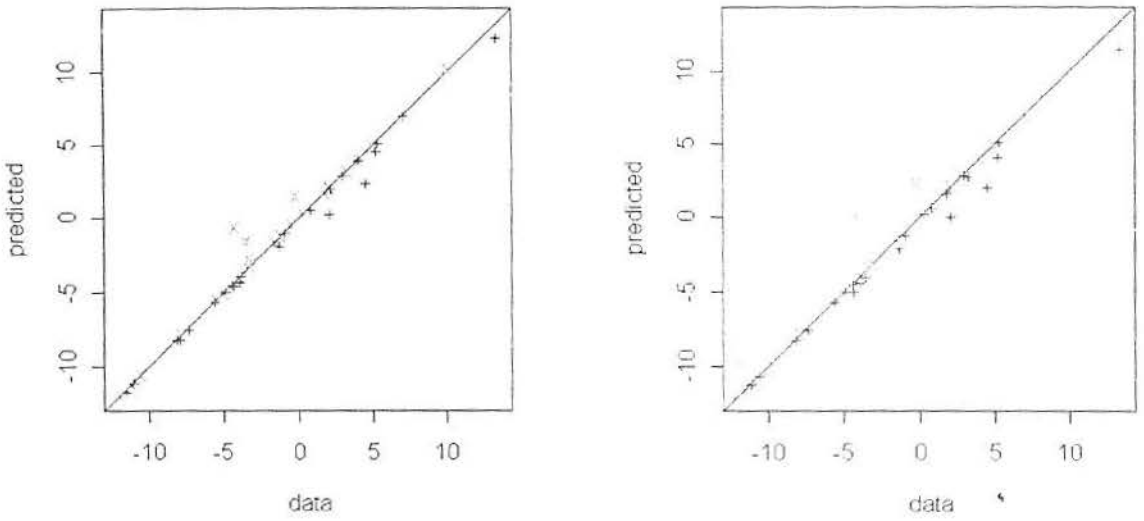

Figura 5.16 - Diagrama de dispersão entre valores observados e preditos pelo modelo ajustado por máxima verossimilhança (esquerda) e por MQP_2 (direita).

Analisando a Figura 4.16, vemos que ambos os modelos podem ser considerados satisfatórios, com mínima vantagem para o modelo ajustado por máxima verossimilhança, já que ambos estimaram com muita precisão os valores já observados.

Este último gráfico serve apenas para mostrar que, em geral, mesmo diferenças bastante significativas entre modelos de estrutura de covariância espacial podem gerar resultados visuais extremamente semelhantes.

6. O PACOTE COMPUTACIONAL GeoR PARA ANÁLISE GEOESTATÍSTICA

Para a realização das técnicas de Geoestatística ao longo desta monografia, utilizou-se o pacote GeoR. O pacote GeoR foi desenvolvido para ser um módulo de análise de dados geoestatísticos utilizando o programa R.

O programa R consiste de um sistema de recursos estatísticos disponível gratuitamente na Internet através do site *http: cran.r-project.org*. Também se encontram disponíveis na Internet inúmeros pacotes de análises estatísticas, desenvolvidos por pesquisadores das mais diversas áreas aplicadas, que utilizam o programa R. Grande parte destes pacotes, inclusive o GeoR, podem ser adquiridos gratuitamente pelo site *http://cran.r-project.org/bin/windows/contrib/PACKAGES'*.

Paralelamente ao GeoR, construiu-se também o pacote GeoS, que funciona no ambiente do software S-PLUS. Ambos GeoR e GeoS foram desenvolvidos por Peter J. Diggle e Paulo J. Ribeiro Jr. quando estes implementavam métodos geoestatísticos sob a perspectiva baseada em modelos, conforme descrito em Diggle & Ribeiro (2000). Maiores detalhes sobre a documentação e utilização destes pacotes podem ser encontrados no site *http://www.maths.lancs.ac.uk/~ribeiro/geoR.html*.

Por se tratar de um programa disponível gratuitamente, o pacote GeoR e o programa R se constituem de importantes ferramentas de análise estatística que se encontram ao alcance de qualquer pesquisador.

6.1. INSTALAÇÃO DO PACOTE GeoR

A instalação do pacote GeoR é extremamente simples e direta. Existem duas maneiras de instalá-lo:

a) Instalação via Internet: Neste caso, vá ao menu PACKAGES e clique na opção *Install Package from CRAN*. Assim, o programa irá se conectar com o site do CRAN (*Comprehensive R archive network*) e instalará o programa desejado dentre uma lista de todos os disponíveis;

b) Instalação local: Para a instalação local, primeiro deve-se realizar o download do pacote através de alguns dos sites citados anteriormente. O pacote está disponível em

versão compactada (*.zip) e deverá ser colocado em uma pasta qualquer do computador. Após isso, vá ao menu PACKAGES e clique na opção *Install Package from local zip file*. Após especificar o caminho até o arquivo, o programa realizará o processo de instalação automaticamente.

Após realizada a instalação, deve-se carregar o pacote a fim de utilizá-lo. Para carregá-lo, vá ao menu PACKAGES e clique com o mouse na opção *Load Package*. Escolha o pacote GeoR e clique OK. Outra forma de carregar o pacote é digitar: library(geoR). Quando o pacote é carregado, o programa exibe a confirmação:

```
geoR: a package for geostatistical analysis in R
geoR is now loaded
```

6.2. ANÁLISE GEOESTATÍSTICA

Nesta seção, apresentaremos alguns comandos básicos do pacote GeoR que são importantes a fim de analisar problemas geoestatísticos.

6.2.1. PREPARAÇÃO DO ARQUIVO DE DADOS

Os arquivos contendo os dados devem ser do tipo texto, preferencialmente com terminações TXT e DAT. O arquivo deve conter três colunas, onde as duas primeiras representam as coordenadas geográficas e a terceira representando os valores da variável em estudo. Os dados devem estar separados por caracteres de tabulação ou por espaços em branco. Exemplo:

quinze.dat

x	*y*	*dados*
0.10	0.20	5.00
0.20	0.30	9.00
0.30	0.50	6.00
0.40	0.60	8.00
0.50	0.70	10.00
0.60	0.20	2.00
0.70	0.90	5.00
0.80	0.40	1.00
0.90	0.40	4.00
1.00	0.50	3.00

x	*y*	*dados*
1.10	0.10	21.00
1.20	0.80	15.00
1.30	0.60	14.00
1.40	0.70	16.00
1.50	0.80	19.00
0.50	1.20	8.00
0.90	1.50	18.00
0.10	0.90	18.00
0.90	0.20	14.00
1.20	1.30	25.00

Posteriormente, vá ao menu FILE e clique na opção *Change Dir*. Deve-se especificar o caminho até o local onde se encontra o arquivo preparado. Então, deve-se utilizar o comando:

```
quinze <- read.geodata("quinze.dat", header=TRUE, coords.col=1:2,
                       data.col=3)
```

Este comando colocará os dados do arquivo *quinze.dat* dentro de um objeto chamado *quinze*. Este objeto guardará as informações do arquivo de dados e irá dividir o arquivo em *coordenadas* e *dados*. O comando *header* informa se nossos dados contém, ou não, um cabeçalho contendo o nome das variáveis. Após a digitação deste comando, podemos verificar se o arquivo foi lido ou se houve algum problema através do comando:

```
print(quinze)
```

6.2.2. ANÁLISE EXPLORATÓRIA DOS DADOS

Para a realização de uma análise preliminar nos dados, podemos utilizar os comandos:

```
plot(quinze)
hist(quinze$data)
points(quinze, cex.min=1, cex.max=2.5, pt.sizes="quartiles",
       col=gray(seq(1,0,l=4)))
points(quinze, cex.min=1, cex.max=2.5, pt.sizes="quartiles")
```

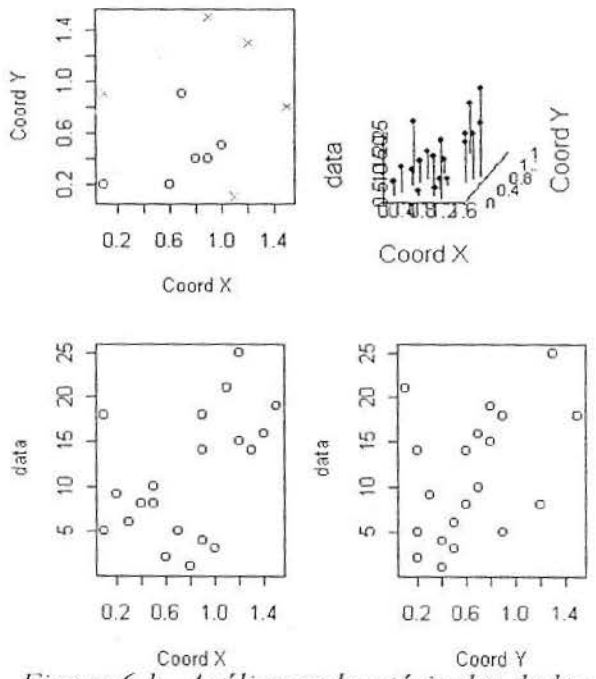

Figura 6.1 - Análise exploratória dos dados.

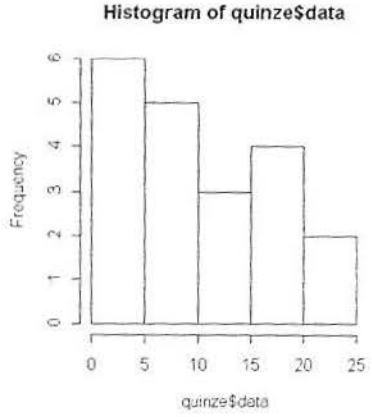

Figura 6.2 - Histograma dos dados.

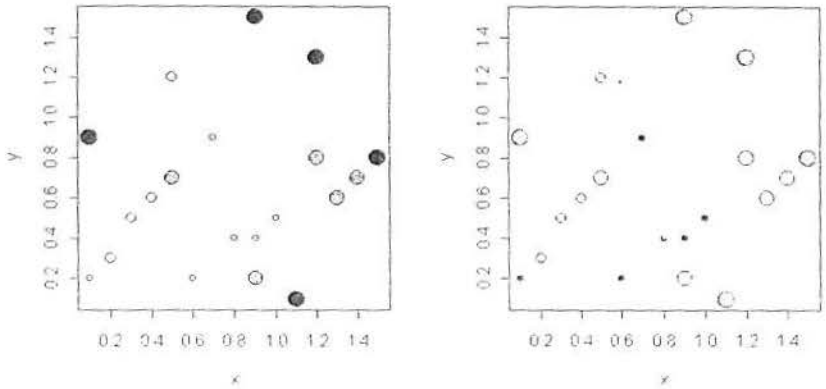

Figura 6.3 - Representação dos valores nos locais amostrados através de duas escalas de cores distintas.

Através destes gráficos, podemos verificar a presença de tendência e valores atípicos.

6.2.3. VARIOGRAMAS AMOSTRAIS

Para o cálculo do variograma empírico e do variograma amostral, utilizamos os comandos:

```
empirico <- variog(quinze, option="cloud", max.dist=1.5)
empirico.robusto <- variog(quinze, option="cloud", max.dist=1.5,
                           estimator.type="modulus")
```

Este último comando calcula o valor do variograma empírico através do estimador robusto. Para visualizarmos os resultados, devemos utilizar os comandos:

```
plot(empirico, main="clássico")
plot(empirico.robusto, main="robusto")
```

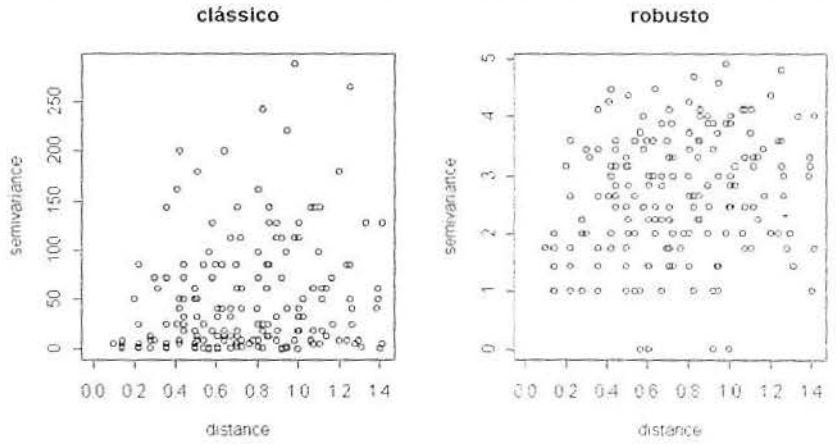

Figura 6.4 - Variograma empírico estimado pelo método clássico (esquerda) e pelo método robusto (direita).

Para o cálculo do variograma amostral, juntamente com os box-plots caracterizando a distribuição do variograma empírico, devemos digitar os comandos:

```
variograma <- variog(quinze, uvec=seq(0,1.5,l=10), bin.cloud=T)
variograma.robusto <- variog(quinze, uvec=seq(0,1.5,l=10),
                      bin.cloud=T, estimator.type="modulus")
```

Os resultados são visualizados através dos comandos:

```
plot(variograma, main="clássico")
plot(variograma.robusto, main="robusto")
plot(variograma, bin.cloud=T, main="clássico")
plot(variograma.robusto, bin.cloud=T, main="robusto")
```

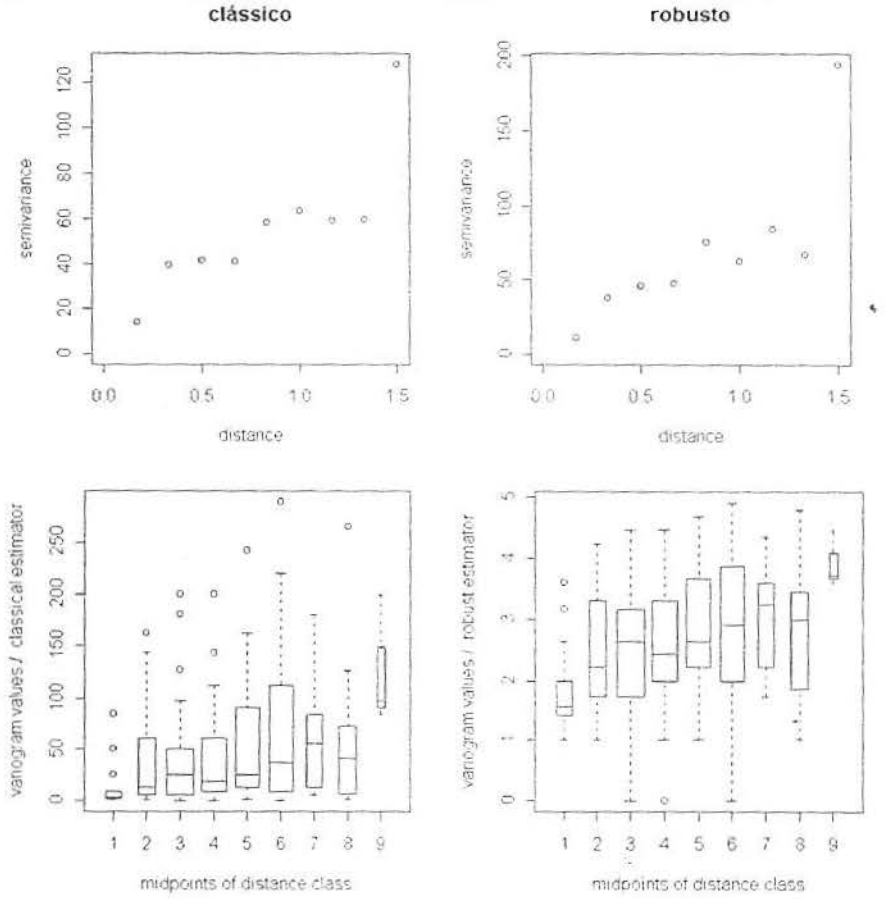

Figura 6.5 - Variograma amostral e box-plots com a distribuição do variograma empírico estimados pelo método clássico (esquerda) e pelo método robusto (direita).

Para o cálculo dos variogramas direcionais, devemos utilizar os comandos:

```
variograma.90 <- variog(quinze, uvec=seq(0,1.5,l=10), direction=pi/2)
variograma.4 <- variog4(quinze, uvec=seq(0,1.5,l=10))
plot(variograma.90, main="Variograma 90º")
plot(variograma.4)
```

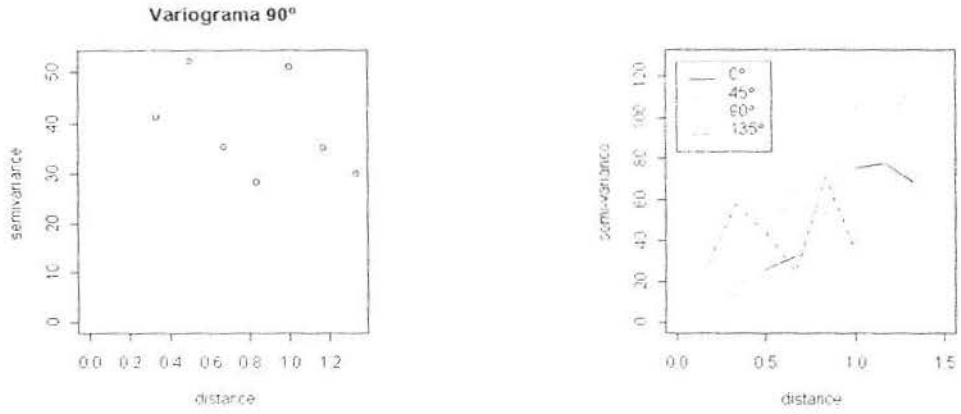

Figura 6.6 - Variograma amostral calculado para a direção de 90° e variograma amostral calculado para as direções de 0°, 45°, 90° e 135°.

Para visualizar melhor o comportamento do variograma amostral, recomenda-se adicionar uma linha ligando os pontos através dos comandos:

```
plot(variograma)
lines(variograma, lty=1, lwd=2, col="red")
legend(0.5, 20, legend="Variograma", lty=1, lwd=2, col="red")
```

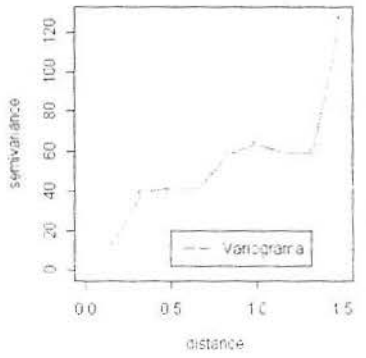

Figura 6.7 - Variograma amostral suavizado.

6.2.4. AJUSTE DE VARIOGRAMAS

Para ajustarmos um modelo de variograma aos nossos dados, devemos primeiro especificar qual critério de ajuste será utilizado:

Mínimos Quadrados Ponderados

O ajuste pode ser feito de três maneiras:

a) Estimando o efeito pepita:

```
mqp <- variofit(variograma, ini=c(80,1.5), cov.model="exponential",
                nugget=0.5, weights="cressie")
```

b) Fixando efeito pepita em zero:

```
mqp.fix.0 <- variofit(variograma, ini=c(80,1.5),
             cov.model="exponential", fix.nugget=T, weights="cressie")
```

c) Fixando o efeito pepita em algum valor arbitrário:

```
mqp.fix <-  variofit(variograma, ini=c(80,1.5),
            cov.model="exponential", fix.nugget=T, nugget=0.15,
            weights="cressie")
```

onde

- *variograma* : objeto que representa o variograma amostral calculado anteriormente
- *ini*: valores dos parâmetros σ^2 e ϕ, respectivamente, que serão considerados como valores iniciais de interação no algoritmo numérico utilizado.
- *cov.model*: identifica qual a família de funções de correlação que será utilizada. Entre as opções estão as funções "Exponential", "Matern", "Gaussian" e "Spherical".
- *nugget*: valor inicial de interação no algoritmo numérico utilizado para o parâmetro τ^2.
- *weights*: ponderação do critério de mínimos quadrados. As opções são "equal" (MQO), "cressie" (MQP_2) e "npairs" (MQP_3).

Máxima Verossimilhança

O ajuste pode ser feito de várias maneiras:

a) Estimando o efeito pepita:

```
mv <- likfit(quinze, ini=c(80,1.5), nug=0.5)
```

b) Fixando efeito pepita em zero:

```
mv.fix.0 <- likfit(quinze, ini=c(80,1.5), fix.nugget=T)
```

c) Fixando o efeito pepita em algum valor arbitrário:

```
mv.fix <- likfit(quinze, ini=c(80,1.5), fix.nugget=T, nugget=0.15)
```

d) Utilizando o critério da máxima verossimilhança restrita:

```
mvr <- likfit(quinze, ini=c(80,1.5), nug=0.5, method="RML")
```

e) Estimando o valor da transformação de Box-Cox:

```
mv.lambda <-  likfit(quinze, ini=c(80,1.5), nug=0.5,
              fix.lambda = FALSE, lambda = 0.5)
```

f) Estimando parâmetros de tendência:

```
mv.tend <- likfit(quinze, ini=c(80,1.5), nug=0.5, trend="1st")
```

onde

- *quinze*: objeto que representa o banco de dados
- *ini*: valores dos parâmetros σ^2 e ϕ, respectivamente, que serão considerados como valores iniciais de interação no algoritmo numérico utilizado.
- *cov.model*: identifica qual a família de funções de correlação que será utilizada. Entre as opções estão as funções "Exponential", "Matern", "Gaussian" e "Spherical".
- *nug*: valor inicial de interação no algoritmo numérico utilizado para o parâmetro τ^2.
- *method*: caso igual a "RML", informa que será utilizado o método da máxima verossimilhança restrita.
- *lambda*: valor inicial de interação no algoritmo numérico utilizado para o parâmetro de transformação λ.
- *trend*: informa o grau do polinômio que representa a média. As opções são "cte" (média constante), "1st" (tendência linear) e "2st" (tendência quadrática).

Após ajustado um modelo, podemos ver os parâmetros estimados através de um dos comandos:

```
summary(mv.tend),  summary(mqp),  summary(mv.lambda), etc.
```

Este comando nos fornece resultados do tipo:

```
Estimation method: maximum likelihood
Parameters of the mean component (trend):
 beta0  beta1  beta2
4.3297 7.6472 4.7808
Parameters of the spatial component:
  correlation function: exponential
    (estimated) variance parameter sigmasq (partial sill) = 34.4117
    (estimated) cor. fct. parameter phi (range parameter) = 0.3428
  anisotropy parameters:
    (fixed) anisotropy angle = 0 ( 0 degrees )
    (fixed) anisotropy ratio = 1
Parameter of the error component:
    (estimated) nugget = 0
Transformation parameter:
    (fixed) Box-Cox parameter = 1 (no transformation)
```

6.2.5. VEROSSIMILHANÇA RELATIVA

Para visualizarmos a função de máxima verossimilhança relativa, devemos utilizar o comando:

```
ver.rel <- proflik(mv, quinze, sill.val=seq(20,150,l=11),
           range.val=seq(0.1,2.5,l=11), uni.only=TRUE)
plot(ver.rel)
```

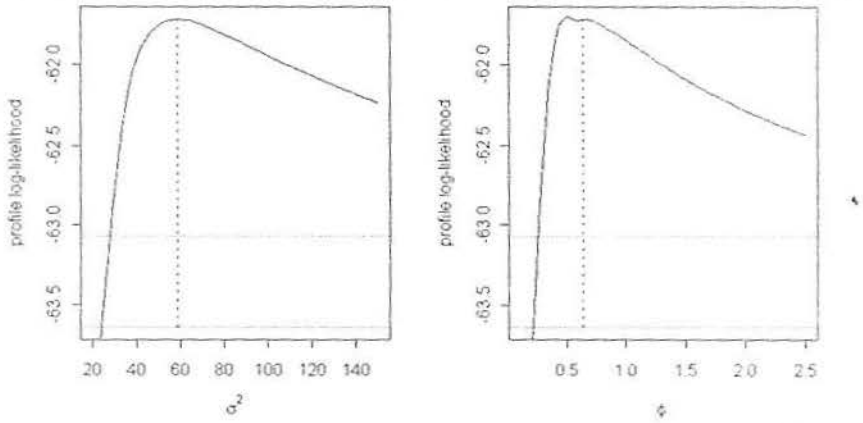

Figura 6.8 - Verossimilhança relativa para os parâmetros σ^2 e ϕ.

onde devemos informar o intervalo (ou sequência) de valores para cada parâmetro desejado. Caso desejarmos um gráfico em duas dimensões, devemos utilizar o comando:

```
ver.rel <- proflik(mv, quinze, sill.val=seq(20,150,l=11),
           range.val=seq(0.1,2.5,l=11), sill.range.val=TRUE,
           nugget.val=seq(0,10,l=11), sillnugget.values=FALSE,
           rangenugget.values=FALSE, uni.only=FALSE, bi.only=TRUE)
plot(ver.rel)
```

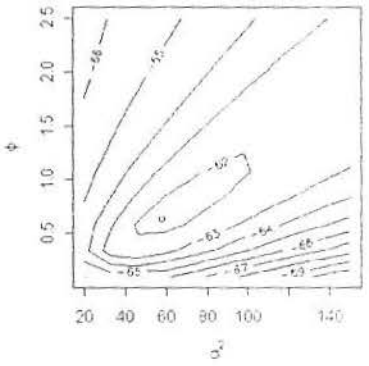

Figura 6.9 - Verossimilhança relativa em duas dimensões para os parâmetros σ^2 e ϕ.

6.2.6. ADICIONANDO O VARIOGRAMA ESTIMADO AO VARIOGRAMA AMOSTRAL

Para adicionar o modelo de variograma estimado ao variograma amostral, deve-se utilizar os seguintes comandos:

```
plot(variograma)
lines(mqp,lty=1,lwd=2,col="blue")
lines(mv,lty=1,lwd=2,col="red")
legend(1,35,legend=c("MQP","MV"),lwd=c(2,2),col=c("blue","red"))
```

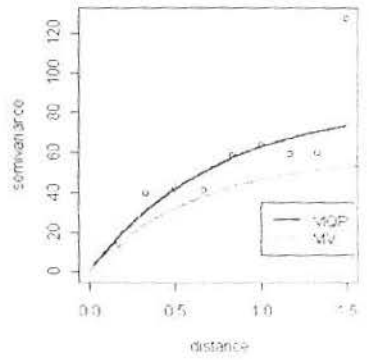

Figura 6.10 - Variogramas ajustados pelos critérios de MQP_2 (em azul) e máxima verossimilhança (em vermelho)

6.2.7. LIMITES SUPERIORES E INFERIORES PARA O VARIOGRAMA AMOSTRAL

Se quisermos calcular os limites (*envelopes*) do variograma, que servem para nos mostrar a variação do variograma amostral para um dado modelo, devemos utilizar o comando:

```
limites <- variog.model.env(quinze, obj.var=variograma, model=mqp,
                            nsim=30)
plot(variograma, envelope=limites)
```

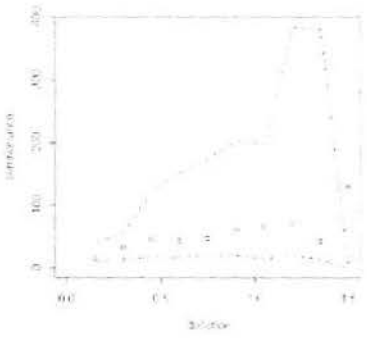

Figura 6.11 - Limites para o variograma amostral.

onde *nsim* informa o número de simulações realizadas. Para a obtenção de simulações, o software R necessita de uma "semente" aleatória. Para especificar uma semente, o programa R fornece uma enorme variedade de métodos. Um exemplo é utilizar o comando:

```
RNGkind("Super")
```

que fornece a semente Super-Duper de Marsaglia. Maiores esclarecimentos sobre os diferentes métodos de introduzir sementes aleatórias podem ser encontrados no manual de referência do software R (*The R Reference Index*, versão 1.3.0, 2001)

6.2.8. PREDIÇÃO ESPACIAL

A fim de realizar a predição espacial, primeiro devemos especificar a malha de pontos que desejamos estimar. Isto é feito através do comando:

```
grid <- expand.grid(seq(0, 1.5, l=30), seq(0, 1.5, l=30))
```

onde as sequências representam os intervalos de coordenadas que fazem parte da malha. O número trinta representa o número de pontos que serão estimados no eixo de coordenadas correspondente. Neste caso, temos 30 vezes 30 pontos na malha a serem preditos.

Para a realização da krigeagem simples, com parâmetros estimados pela máxima verossimilhança, devemos utilizar o comando:

```
krigeagem <- krige.conv(quinze, loc=malha,
             krige=krige.control(type.krige="SK", trend.d="cte",
             beta =15.14, cov.pars=mv$cov.pars))
```

onde *trend.d* indica o grau do polinômio que representa a média e *beta* representa o valor dos parâmetros da média. Neste caso, temos média constante com parâmetro β_0 estimado por máxima verossimilhança e valor igual a 15,14. O valor "SK" no comando *type.krige* representa o tipo de krigeagem a ser realizada. As opções são SK (krigeagem simples), OK (krigeagem ordinária) e UK (krigeagem universal).

Se desejamos realizar a krigeagem supondo que todos os parâmetros já são conhecidos, podemos fixar os parâmetros através do comando:

```
krigeagem.fixa <-  krige.conv(quinze, loc=malha,
                   krige=krige.control(type.krige="SK", trend.d="cte",
                   beta =15.14, cov.pars=c(58,0.63), nugget=0))
```

Para visualizar o mapa da região com os resultados da krigeagem, utilizamos o comando:

```
image(krigeagem, loc=malha, coords=quinze$coords,
      col=heat.colors(256), x.leg=c(1.4,1.49), y.leg=c(1.3,1.55),
      cex.leg=0.7,vertical=TRUE)
```

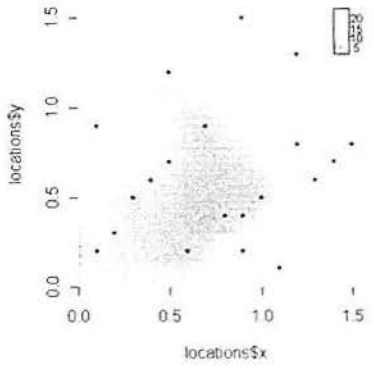

Figura 6.12 - Valores preditos por krigeagem simples.

onde *x.leg* e *y.leg* representam as coordenadas da legenda e *cex.leg* o tamanho da fonte . O comando *col* identifica a escala de cores da imagem. Entre outras, temos a escala de cores "*terrain.colors*(nº de cores)", "*topo.colors*(nº de cores)", "*cm.colors*(nº de cores)" e "*rainbow*(nº de cores)".

Para visualizar os valores da variância de predição, deve-se utilizar o comando:

```
image(krigeagem, loc=malha, coords=quinze$coords,
   values=krigagem$krige.var, col=heat.colors(256), x.leg=c(1.4,1.49),
   y.leg=c(1.3,1.55), cex.leg=0.7, vertical=TRUE)
```

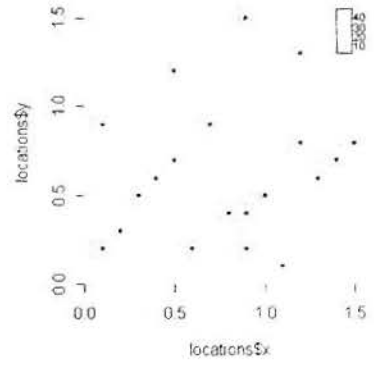

Figura 6.13 - Variâncias de predição.

Outra forma de visualizar os valores preditos pela krigeagem é através de superfícies. Para isto, deve-se utilizar o comando:

```
persp.kriging(krigeagem, loc=malha, values=krigagem$predict, theta=0,
              col=heat.colors(256))
persp.kriging(krigeagem, loc=malha, values=krigagem$predict, theta=45,
              col=heat.colors(256))
```

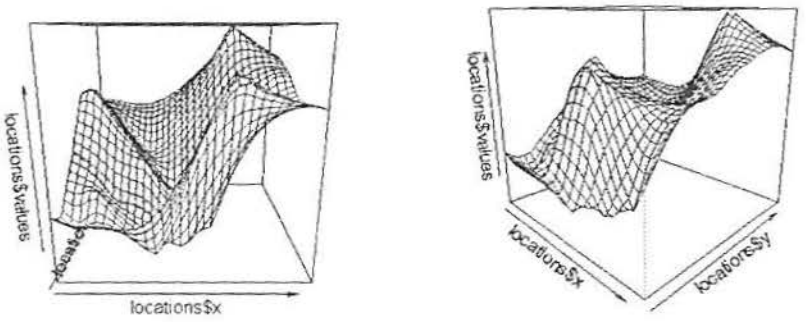

Figura 6.14 - Superfície com os valores preditos visualizada por ângulos θ diferentes.

6.2.9. VALIDAÇÃO

Para compararmos os valores preditos com os valores observados, utilizamos o seguinte comando:

```
valid <- xvalid(quinze, model=mv)
plot(valid, quinze)
```

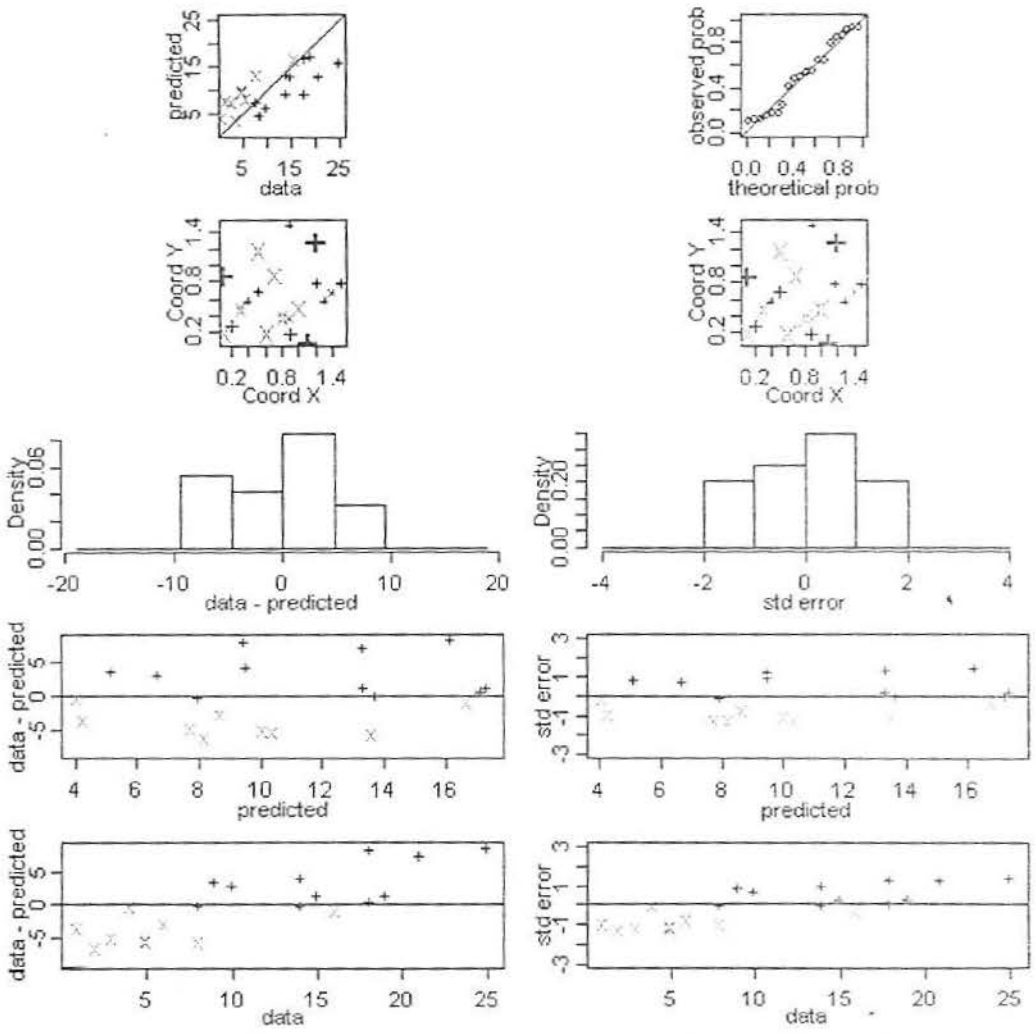

Figura 6.15 - Comparação entre valores preditos e observados, juntamente com análise dos resíduos de predição.

6.2.10. SIMULAÇÃO GAUSSIANA

Após a definição de uma semente aleatória, conforme descrito na seção 6.2.7, podemos gerar um processo Gaussiano através do seguinte comando:

```
sim <-  grf(80, grid="irreg", cov.model="spherical",
            cov.pars = c(1, .25), nugget=0)
plot(sim)
points(sim)
```

onde *80* representa o número de pontos simulados e *grid* informa o tipo de malha amostral. As opções de malha são "irreg" (irregular) e "reg" (regular).

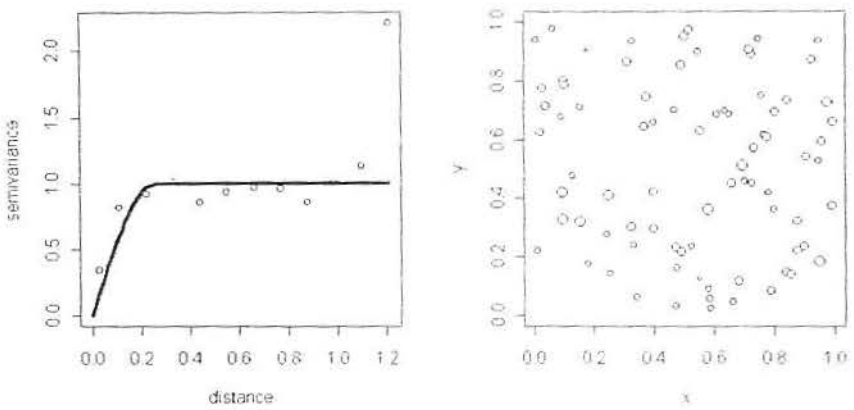

Figura 6.16 - Variograma amostral juntamente com o verdadeiro modelo de variograma (esquerda) e visualização dos locais amostrados para o processo Gaussiano simulado (direita).

7. CONSIDERAÇÕES FINAIS

A Geoestatística tem sido amplamente utilizada por profissionais e pesquisadores das mais variadas áreas da ciência. Infelizmente, por questão de simplicidade — e muitas vezes desconhecimento — as técnicas de Geoestatística têm sido aplicadas sem a preocupação com os efeitos e problemas da não especificação de um modelo apropriado ao fenômeno em questão.

Uma das principais razões que colaboram com esta prática é o fato de que muitas vezes os resultados de predição espacial para um certo fenômeno, obtidos a partir da utilização de modelos diferentes, fornecem estimativas bastante semelhantes. Este tipo de prática nos fornece um erro adicional semelhante ao erro existente quando, na estimação da média de uma distribuição Normal, tratamos a variância amostral como o verdadeiro valor da variância populacional. Procedendo desta forma, utilizaríamos o valor de 1,96 para construir um intervalo com 95% de confiança para a média populacional ao invés de utilizar o valor corresponde da distribuição t de Student. Apesar dos valores serem semelhantes, para amostras pequenas as diferenças podem ser bastante significativas. Neste caso particular, é possível quantificar o erro existente ao realizar tal prática. Entretanto, na maior parte dos casos, desconhecemos a magnitude deste erro.

Apesar deste fato, não podemos diminuir a importância das técnicas de Geoestatística que não necessitam de suposições à respeito da forma ou distribuição do processo gerador dos dados. Para muitos fenômenos estudados, uma simples interpolação de valores pode fornecer os resultados esperados de forma bastante simples. Entretanto, a utilização de técnicas de Geoestatística — ou de qualquer área da Estatística — que são baseadas em um modelo especificado de forma coerente, normalmente produzem resultados ótimos. O fato de não conseguirmos ajustar algum modelo conhecido aos nossos dados deve funcionar como adicional motivação para buscarmos novas técnicas que possam se adequar ao nosso problema.

O avanço dos recursos computacionais, juntamente com a utilização de métodos de simulação de Monte Carlo, tornam possível análises baseadas em modelos cada vez mais complexos que cobrem uma vastidão de fenômenos do conhecimento humano. Em Geoestatística, estes avanços tem sido importantes para o estudo de

fenômenos onde o processo gerador dos dados não é Gaussiano e, consequentemente, é necessário o uso de técnicas de predição baseadas em estimadores não-lineares. Obras de referência nesta área são Diggle, Moyeed e Tawn (1998) e Diggle & Ribeiro (2000).

Além da importância da escolha do modelo adequado, salientamos que uma análise exploratória antes de qualquer tentativa de ajustar modelos de covariância espacial é de vital importância. A detecção a priori de média não-constante, presença de valores atípicos e anisotropia nos nossos dados poupará tempo e tornará a fase de ajuste de modelos de covariância espacial uma tarefa significativamente mais simples.

8. REFERÊNCIAS BIBLIOGRÁFICAS

Bailey, T.C. and Gatrell, A.C. (1995). *Interactive spatial data analysis*. Essex: Longman Scientific & Technical.

Costa Neto, P.L.O. (1977). *Estatística*. São Paulo: Ed. Edgard Blücher Ltda.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: Wiley.

Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-Based Geostatistics. *Applied Statistics*, 47, 299-350.

Diggle, P.J., Ribeiro Jr, P.J. (2000). *Model Based Geostatistics*. 14° SINAPE, Associação Brasileira de Estatística - ABE.

Gamerman, D. e Migon, H.S.(1993). *Inferência Estatística: Uma Abordagem Integrada*. Universidade Federal do Rio de Janeiro, Textos de Métodos Matemáticos, n°27.

Harvey, AC (1981). *Times Series Models*. Oxford: Philip Allan

Journel, A.G. e Huijbregts, C.J. (1978). *Mining Geostatistics*. London:-Academic Press.

Meyer, P.L.(1983). *Probabilidade: Aplicações à Estatística*. 2ª Edição. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. (1986). *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Ed. Cambridge University Press.

The R Reference Index. *Manual de referência do software R*, versão 1.3.0. Atualizado em 2001

ANEXOS

ANEXO 1 - PARÂMETROS ESTIMADOS PARA OS MODELOS SIMULADOS DA SEÇÃO 4.7

i) Modelo Esférico: $\theta = (\sigma^2 = 0{,}1;\ \ \phi = 0{,}1;\ \ \tau^2 = 0)$

$n = 50$

Mínimos Quadrados Ponderados (2)
variância = 0.1001
ϕ = 0.0273
pepita = 0

Máxima Verossimilhança
Beta (média) = 0.0372
variância = 0.05
ϕ = 0
pepita = 0.0483

$n = 100$

Mínimos Quadrados Ponderados (2)
variância = 0.0754
ϕ = 0.1099
pepita = 0

Máxima Verossimilhança
Beta (média) = 0.0024
variância = 0.0753
ϕ = 0.0975
pepita = 6e-04

ii) Modelo Esférico: $\theta = (\sigma^2 = 100;\ \ \phi = 0{,}4;\ \ \tau^2 = 20)$

$n = 50$

Mínimos Quadrados Ponderados (2)
variância = 140.0158
ϕ = 0.4656
pepita = 0

Máxima Verossimilhança
Beta (média) = -2.398
variância = 106.1822
ϕ = 0.3611
pepita = 9.9935

$n = 100$

Mínimos Quadrados Ponderados (2)
variância = 106.1741
ϕ = 0.3475
pepita = 23.8098

Máxima Verossimilhança
Beta (média) = 0.7917
variância = 108.8559
ϕ = 0.4215
pepita = 20.6282

iii)Modelo Exponencial Potência $k = 2$: $\theta = (\sigma^2 = 50;\ \phi = 0{,}3;\ \tau^2 = 0)$

$n = 50$

Mínimos Quadrados Ponderados (2)
variância = 40.8847
ϕ = 0.2381
pepita = 0

Máxima Verossimilhança
Beta (média) = -2.2341
variância = 51.1484
ϕ = 0.3065
pepita = 0

$n = 100$

Mínimos Quadrados Ponderados (2)
variância = 31.2959
ϕ = 0.2205
pepita = 0

Máxima Verossimilhança
Beta (média) = -1.4627
variância = 45.4373
ϕ = 0.2981
pepita = 0

iv)Modelo Exponencial Potência com k = 1: $\theta = (\sigma^2 = 40;\ \ \phi = 1;\ \ \tau^2 = 5)$

$n = 50$

Mínimos Quadrados Ponderados (2)
variância = 69.5401
ϕ = 1.5486
pepita = 3.6085

Máxima Verossimilhança
Beta (média) = -0.9137
variância = 23.6498
ϕ = 0.2563
pepita = 1.2803

$n = 100$

Mínimos Quadrados Ponderados (2)
variância = 23.0619
ϕ = 0.2201
pepita = 0

Máxima Verossimilhança
Beta (média) = 1.7421
variância = 15.8599
ϕ = 0.3983
pepita = 3.5104

- 12 estações no circulo sobre o raio de 300m
- 12 estações no circulo sobre o raio de 500m
- 06 estações no circulo sobre o raio de 2500m (estações de referência)

Os dados utilizados nesta monografia são provenientes dos pontos de amostragem de sedimentos formando um total de 47 pontos. O motivo da redução da amostra se deve ao fato de não consideramos as amostras obtidas nas estações de referência e pelo fato da impossibilidade da coleta de material em um dos locais amostrados.

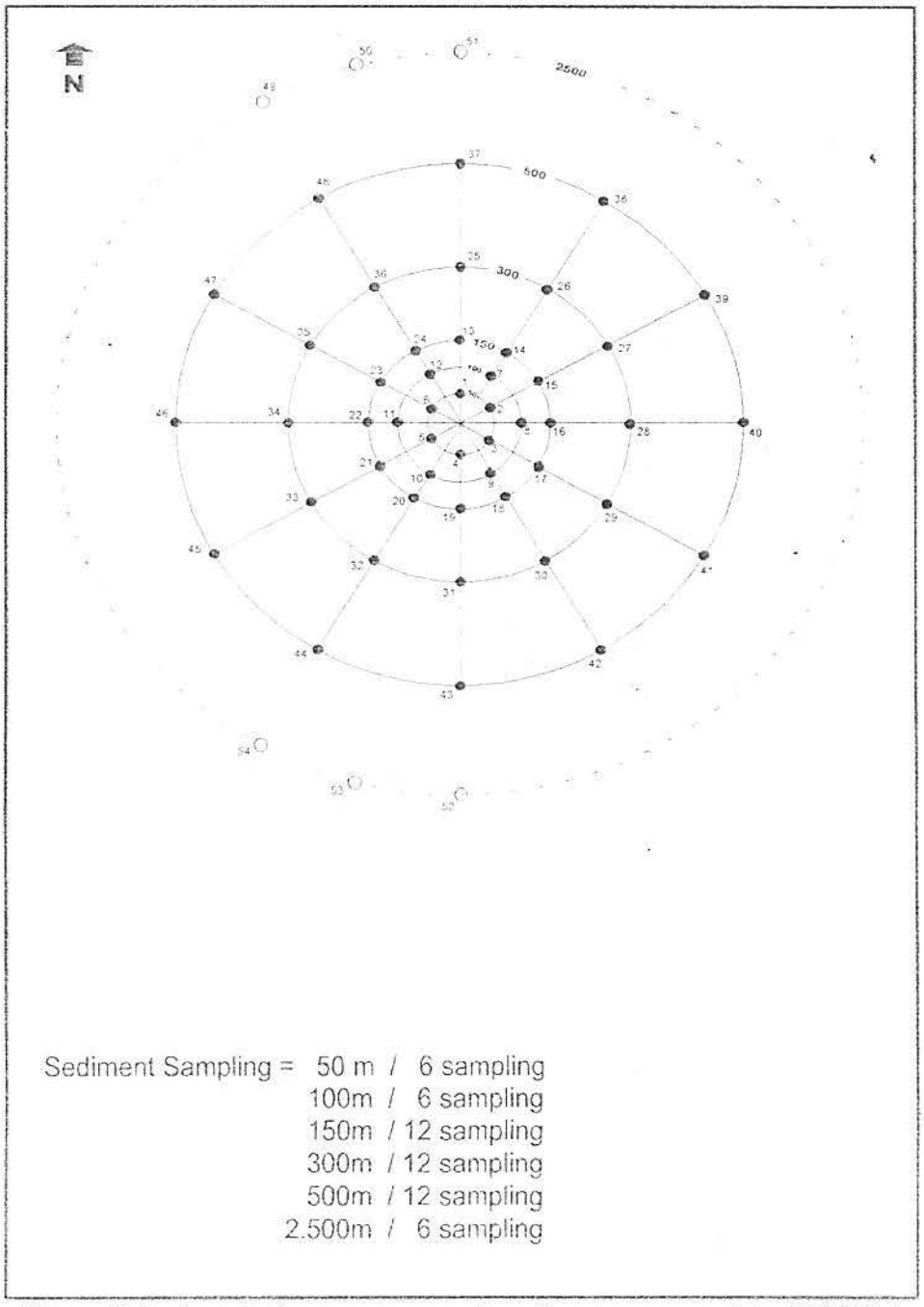

Figura 7.1 - Locais de amostragem de sedimentos no projeto MAPEM.

ANEXO 2 - O PROJETO MAPEM

O projeto de *Monitoramento Ambiental em Atividades de Perfuração Marítima* (MAPEM) é um Projeto FINEP/IBP, com a coordenação do Instituto de Geociências da UFRGS e com a co-participação do Instituto de Matemática, Instituto de Química e Instituto de Informática da UFRGS, juntamente com o Núcleo de Estudos Marítimos (NEMAR) da UFSC.

O projeto MAPEM ainda se encontra em andamento e tem como principal objetivo avaliar os efeitos dos fluidos Não-Aquosos (NAF) associados com o descarte de cascalho, em dois locais de perfuração (um localizado em águas superficiais e outro localizado em águas profundas) na Bacia de Campos, Brasil. Outro objetivo do programa é determinar o grau de impacto ambiental após a atividade de perfuração e o grau de recuperação do local um ano após a perfuração. Este programa foi estruturado para obter uma coleta abrangente de dados nos dois locais, que poderão ser usados no futuro para orientar o desenvolvimento de estratégias de monitoramento, baseados em dados científicos.

Dentre os objetivos relacionados com a Geoestatística, podemos destacar que o projeto procura determinar se existe algum impacto resultante das operações de perfuração nos locais amostrados através de um controle da variação natural no local.

Esta monografia utilizou os dados obtidos na primeira fase de coleta de dados, onde foram obtidos dados das seguintes áreas:

- Biologia (macrofauna, meiofauna).
- Química (Hidrocarbonetos, metais).
- Geologia (Granulometria, *TOC*)

Os dados referentes a esta primeira coleta foram obtidos antes da perfuração dos poços e, por este motivo, objetivam servir de fonte para o estudo variação natural da região.

Os dados utilizados desta primeira coleta se constituem de cinqüenta e quatro (54) pontos de amostragem em cada local de perfuração de acordo com o modelo de amostragem (ver Figura 7.1) definido abaixo:

- 06 estações no círculo sobre o raio de 50m
- 06 estações no círculo sobre o raio de 100m
- 12 estações no círculo sobre o raio de 150m