Resumo

A preocupação em identificar os parâmetros do reservatório que mais interferem no escoamento de fluidos e modelá-los numa escala compatível com o estudo de simulação de fluxo é antiga. Desde a década de 70 sentiu-se a necessidade de descrever adequadamente esses parâmetros e vários modelos e escalas de heterogeneidade foram propostos com o objetivo de defini-las, como também de definir as incertezas inerentes ao conhecimento do reservatório. Dentre as técnicas de obtenção de dados, na fase de perfuração de um poço, podemos destacar a perfilagem, pelo seu baixo custo em relação às demais técnicas. Por meio dela, é obtida uma série de dados dos poços, chamados perfis, cuja interpretação permite uma avaliação da formação geológica em estudo, ou seja, da jazida petrolífera. A geoestatística enfatiza o contexto geológico em que os dados foram obtidos, a relação espacial entre esses dados e os valores medidos em diferentes suportes volumétricos e graus de precisão. Modelos da geoestatística baseiam-se, em parte, na teoria das probabilidades, reconhecendo e incorporando as incertezas advindas do processo de obtenção dos dados. A modelagem torna-se uma ferramenta importante, pois permite simular as regiões de onde se possui pouco ou nenhum conhecimento.

Abstract

The concern in identifying the parameters of the reservoir that are relevant in fluid draining and shape them in a compatible scale with the study of flow simulation is old. Since 70 years it was felt necessity to describe adequately these parameters and some models and scales of heterogeneity had been considered with the objective to define heterogeneities, as also the uncertainties inherent to the knowledge of the reservoir. Among the sampling techniques in the well perforation stage, we can detach the logging, because its low cost in relation to others techniques. Through it, are gotten a series of data about the well, called profiles, whose interpretation allows an evaluation of the geologic formation in study. The geostatistic emphasizes the geologic context where the data had been gotten, the spatial relation among these data and the values measured in different volumetric supports and precision degrees. Models of geostatistic are based, in part, in probability theory, recognizing and incorporating the happened uncertainties of the process of attainment of the data. The modeling becomes an important tool, therefore it allows to simulate the regions where little or any information is available.

Universidade Federal de Campina Grande
Centro de Ciências e Teconologia
Programa de Pós-Graduação em Matemática
Curso de Mestrado em Matemática

A Geoestatística Aplicada à Avaliação e Caracterização de Reservatórios Petrolíferos

por

Ana Cristina Brandão da Rocha

sob orientação do

Prof. Dr. Francisco Antônio Morais de Souza

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, como requisito parcial para obtenção do título de Mestre em Matemática.

Campina Grande - PB

Outubro/2005

A Geoestatística Aplicada à Avaliação e Caracterização de Reservatórios Petrolíferos

por

Ana Cristina Brandão da Rocha

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, como requisito parcial para obtenção do título de Mestre em Matemática.

Área de Concentração: Matemática

Aprovada por:

Prof. Dr. Manoel Raimundo de Sena Junior

Prof. Dr. Antonio José da Silva

Prof. Dr. Francisco Antônio Morais de Souza

Orientador

Universidade Federal de Campina Grande

Centro de Ciências e Tecnologia

Programa de Pós-Graduação em Matemática

Curso de Mestrado em Matemática

Outubro/2005

Agradecimentos

Primeiramente, agradeço a Deus, pela Sua infinita misericórdia e por ter me dado forças para superar os obstáculos desta caminhada; e à Maria, que nunca cessou de interceder por mim em todos os momentos;

Aos meus pais, Antonio e Terezinha, pelo grande esforço, pela credibilidade, pela torcida e pelas orações que fizeram para que eu chegasse até aqui;

Ao meu irmão Alysson (Lasson), pela facilidade que tem de me fazer sorrir, ajudando a quebrar tantas barreiras;

A Célio, que com carinho sempre me deu força, dividiu comigo as alegrias e as preocupações e me ajudou em tudo;

Ao professor Francisco Morais, pela orientação, paciência, dedicação, apoio, amizade e por tanto me inspirar;

À ANP (Agência Nacional do Petróleo, Gás e Biocombustíveis) e à FINEP (Financiadora de Estudos e Projetos), pela concessão da bolsa e pelo apoio recebido;

Aos professores Manoel Raimundo de Sena Junior e Antonio José da Silva, por me avaliarem;

A todos os professores do DME/UFCG, entre eles os professores Marco Aurélio Soares Souto, Antonio José da Silva e Aparecido Jesuino de Souza, e aos professores José Edilson Neves, pesquisador visitante do PRH-25/ANP/MCT, Ramdayal Swarnakar e Severino Rodrigues de Farias Neto, do DEQ/UFCG, pelas disciplinas que lecionaram e que tanto contribuíram para minha formação;

A Lindomberg que com tanta paciência me ajudou com os programas para a apresentação do trabalho;

A todos os funcionários do DME/UFCG, entre eles Salete, Valdir (que furou a greve no feriadão de Outubro para me ajudar), Marcelino, Dona Argentina e Du;

Aos amigos Iraponil, Jesus Robson, Juliana, Tatiana, Jesualdo, Rosângela, Areli, Grayci, Lya, Marta, Jacqueline, Jackelya, Hallyson, Lino, Thiciany, Jefferson, Lauri-

clécio, Érica, Danielle, Magno (e sua esposa Gorete), Fabriciano, Ricardo, entre outros; Sejam eles alunos do curso de Matemática ou não, alunos do mestrado em Matemática ou não, da minha turma ou não, companheiros nas disciplinas ou não, companheiros de estudo em dias de sábado ou não, irmãos por parte de orientador ou não. Gostaria de agradecer a todos por compartilhar comigo cada momento, e compreender minhas ausências... Muito obrigada por tudo, principalmente pela amizade!

Por fim, agradeço a todos que diretamente ou indiretamente contribuíram para a realização deste trabalho. Muito obrigada! Só Deus poderá recompensar-vos!

Dedicatória

Aos meus pais, Antônio e Terezinha.

Conteúdo

Introdução

As atividades e estudos que visam determinar, em termos qualitativos e quantitativos, o potencial de uma jazida petrolífera, isto é, a sua capacidade produtiva e a valoração de suas reservas de óleo e gás, é um fator de grande interesse da indústria petrolífera.

Quando uma jazida petrolífera é descoberta, vários procedimentos são desenvolvidos com a finalidade de levantar informações, visando um melhor entendimento dessa jazida. Assim, desde a exploração de um campo até o seu abandono, uma enorme quantidade de dados é coletada. Dentre as técnicas de obtenção de dados na fase de perfuração de um poço, podemos destacar a perfilagem, pelo seu baixo custo em relação às demais técnicas. Através dela, são obtidos inúmeros dados dos poços, denominados *dados de perfis*, cuja interpretação permite uma avaliação da formação em intervalos maiores e em condições reais do poço.

A indústria petrolífera tem sido motivada a utilizar renovadas técnicas de caracterização de reservatório, pelos altos investimentos necessários no desenvolvimento de campos heterogêneos e pelo desejo e necessidade de aumentar o fator de recuperação final das jazidas. Dentre essas técnicas, podemos citar as técnicas geoestatísticas, subárea da estatística, com aplicação crescente, especialmente quando há a utilzação de dados sísmicos tridimensionais.

Muitas são as vantagens da aplicação da Geoestatística. Uma delas é o fato de ela necessitar da interdisciplinaridade, assegurando uma maior troca de informações entre geólogos, engenheiros de petróleo e estatísticos e uma melhor interpretação da realidade geológica em estudo.

Este trabalho tem como objetivo fazer uma análise Geoestatística de dados dos

perfis DT, GR, ILD, NPHI e RHOB, que estão descritos no Capítulo 1, relativos a poços petrolíferos do Campo Escola de Namorado, situado na bacia de Campos, no estado do Rio de Janeiro.

No Capítulo 1, faremos uma breve explanação a respeito da Avaliação das Formações, etapa de fundamental importância no desenvolvimento de um campo petrolífero.

No Capítulo 2, abordaremos as Técnicas Multivariadas, cuja aplicação em estudos com dados geológicos é bastante ampla.

No Capítulo 3, tratamos das técnicas da Geoestatística, enfatizando suas principais vantagens em ralação às técnicas estatísticas na avaliação e caracterização de reservatórios.

No Capítulo 4, apresentaremos os resultados obtidos neste trabalho.

Capítulo 1

Avaliação de Formações

Denomina-se Avaliação de Formações as atividades e estudos que visam determinar, em termos qualitativos e quantitativos, o potencial de uma jazida petrolífera, isto é, a sua capacidade produtiva e a valoração de suas reservas de óleo e gás [8]. O dimensionamento das reservas constitui uma etapa muito importante no processo de exploração de um reservatório, visto que estabelece a viabilidade do desenvolvimento do campo de petróleo. Uma das principais ferramentas utilizadas durante esta atividade é a perfilagem a poço aberto, através do teste de formação a poço aberto.

A primeira etapa realizada no desenvolvimento de um campo é a perfuração do poço pioneiro, cuja locação é previamente estabelecida através de estudos geológicos, que propiciam uma dedução da estrutura, da composição e da história geológica da superfície terrestre analisada, e geofísicos, os quais envolvem o estudo das partes profundas da terra que não podem ser vistas, e que têm suas propriedades físicas medidas por meio de sofisticados e apropriados instrumentos, geralmente colocados na superfície.

A existência de acumulações de petróleo depende das características e do arranjo de certos tipos de rochas sedimentares no subsolo. Basicamente, é preciso que existam rochas geradoras que contenham a matéria-prima que se transforma em petróleo e rochas-reservatório, ou seja, aquelas que possuem espaços vazios, chamados poros, capazes de armazenar o fluido. Essas rochas são envolvidas em armadilhas chamadas *trapas*, compartimentos isolados no subsolo onde o fluido se acumula e de onde não

tem condições de escapar. A ausência de qualquer um desses elementos impossibilita a existência de uma acumulação petrolífera. Sendo assim, a existência de uma bacia sedimentar não garante, por si só, a presença de jazidas de petróleo [10].

Ao término da perfuração de um poço, algumas informações muito importantes a respeito das formações por ele atravessadas podem ser obtidas, através da chamada perfilagem final. Dentre as informações, podemos citar: litologia (tipo de rocha), porosidade, espessura, prováveis fluidos existentes nos poros e suas respectivas saturações. Os dados que fornecem essas informações são chamados de *perfis*. É a partir de suas análises que se pode observar quais faixas de profundidade do poço são de interesse econômico potencial para se executar testes de custo bem mais elevado, chamados de testes de formação. Caso não haja faixas de profundidade de interesse, o poço é abandonado.

Usados principalmente na prospecção de petróleo e de água subterrânea, os perfis de poços têm sempre como principal objetivo a determinação da profundidade e a estimativa do volume dos aqüíferos ou da jazida de hidrocarbonetos. Eles são geralmente obtidos por meio de diversos sensores que, acoplados a sofisticados aparelhos eletrônicos, são introduzidos no poço e registram, a cada profundidade, as diversas informações relativas às características físicas das rochas e dos fluidos em seus poros [25].

Existem muitos tipos de perfis, com aplicações as mais diversas, todos com o objetivo de melhor avaliar as formações geológicas quanto à definição de camadas potencialmente produtoras e de proporcionar uma análise detalhada e precisa da composição do fluido das rochas do subsolo. Dentre eles, os principais são:

- **Potencial Expontâneo (SP):** Potencial elétrico naturalmente desenvolvido nas camadas permoporosas, devido à diferença de salinidade que existe entre o fluido de perfuração e a água da formação. Permite uma indicação qualitativa da permeabilidade das rochas, característica imprescindível para o escoamento dos fluidos nela existentes, e o cálculo da resistividade, da salinidade da água da formação e da argilosidade das rochas. Auxilia também na correlação com poços vizinhos [24].

- **Sônico (DT):** É utilizado para obter estimativas das porosidades total e efetiva das rochas, correlação poço a poço, cálculo da velocidade compressional e das

constantes elásticas das rochas, detecção de fraturas e apoio à sísmica para a elaboração do sismograma sintético.

- **Raios Gama (GR):** Detecta a radioatividade natural das rochas, em função dos isótopos do Urânio, Tório e Potássio-40, presentes na natureza. É utilizado para a identificação da litologia, para a identificação de minerais radioativos e para o cálculo percentual da argilosidade. Permite observação da variação granulométrica das camadas e da correlação entre poços.

- **Indução (ILD):** Este perfil fornece leitura aproximada da resistividade total, por meio da medição de propagação de ondas eletromagnéticas e de leituras diagnósticas, onde existem trocas iônicas na superfície de grãos metálicos, tal como acontece em sulfitos. É empregado pela engenharia na diferenciação litológica e em estudos de salinidade de lençóis de água subterrânea.

- **Neutrônico (NPHI):** Os perfis mais antigos medem a quantidade de raios gama de captura, após excitação artificial por meio de bombardeio dirigido de nêutrons rápidos. Os mais modernos medem a quantidade de nêutrons epitermais e/ou termais da rocha, após o bombardeio. São utilizados para estimativas de porosidade, litologia e detecção de hidrocarbonetos leves ou gás. Desta maneira, quanto maior o valor do NPHI, maior a quantidade de poros e, conseqüentemente, maior a probabilidade do fluido ali existente ser extraído.

- **Densidade (RHOB):** Detecta os raios gama defletidos pelos elétrons orbitais dos elementos componentes das rochas, após terem sido emitidos por uma fonte colimada situada dentro do poço. Este tipo de perfil fornece densidade das camadas e permite o cálculo da porosidade e a identificação das zonas de gás. É utilizado também como apoio à sísmica, no cálculo do sismograma sintético [20].

Um estudo mais detalhado sobre a avaliação das formações pode ser encontrado em [16].

Capítulo 2

Técnicas Multivariadas

2.1 Introdução

As técnicas estatísticas multivariadas são bastante utilizadas nas ciências biológicas, na antropologia, na psicologia, nos estudos da ciência política e da economia, na geografia e em outras áreas, pois permitem ao pesquisador a análise simultânea de um grande número de variáveis. Têm-se, em muitos procedimentos, obtido resultados satisfatórios, especialmente falando em estudos relacionados com a geologia, visto que a maioria dos problemas dessa área envolve forças complexas e que interagem umas com as outras, impossibilitando seu estudo separadamente.

Por este motivo, faremos a seguir uma breve descrição a respeito de alguns métodos que são utilizados como ferramenta para a análise de dados geológicos.

2.2 Análise de Conglomerados

Análise de Conglomerados (também denominada Análise de Agrupamentos ou Análise de *Clusters*) é o nome dado às técnicas de análise que dividem os dados em grupos. Estes grupos podem ser constituídos por observações individuais multivariadas ou agrupamentos multivariados de variáveis.

A Análise de Conglomerados classifica objetos e pessoas sem preconceitos, isto é, observando apenas as semelhanças ou dissimilaridades entre elas, sem definir critérios de inclusão prévia em qualquer agrupamento. Os métodos de Análise de Conglomera-

dos tentam organizar um conjunto de indivíduos, para os quais é conhecida informação detalhada, em grupos relativamente homogêneos (*clusters*).

A Análise de Conglomerados tem como base um conjunto de n indivíduos para os quais existe informação sob a forma de p variáveis. O método faz o agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes entre si do que a indivíduos dos demais grupos [15].

Na análise de agrupamentos, é fundamental ter particular cuidado na seleção das variáveis de partida que vão caracterizar cada indivíduo ou caso, e determinar, em última instância, qual o grupo em que deve ser inscrito. Nesta análise não existe qualquer tipo de dependência entre as variáveis, isto é, os grupos configuram-se por si mesmo sem necessidade de ser definida uma relação causal entre as variáveis utilizadas. O método é exploratório e a idéia é, sobretudo, gerar hipóteses, mais do que testá-las, pelo que é necessária a validação posterior dos resultados encontrados através da aplicação de outros métodos estatísticos.

Uma dificuldade inicial é a de não existir uma única via de definição de grupos, isto é, um único critério de partição e/ou agrupamento dos indivíduos (poços) ou casos com base numa única medida de semelhança. Em todos eles se pretende que os grupos sejam coerentes e que se distinguam de maneira significativa uns dos outros. Genericamente, a análise de agrupamentos compreende seis etapas:

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. A escolha da técnica de análise;
4. A definição de uma medida de semelhança ou distância entre cada dois indivíduos;
5. A escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição de um algoritmo de partição/ classificação;
6. A validação dos resultados encontrados.

2.2.1 Seleção das Variáveis

A seleção das variáveis comporta um duplo problema: um problema substantivo, que terá de ser resolvido com o conhecimento prévio do investigador sobre o assunto a estudar e que lhe permitirá escolher dentre os dados disponíveis quais os mais significativos na abordagem do problema, e um outro, de ordem mais estatística que tem a ver com o tipo de variáveis utilizadas, sobretudo quando estas estão definidas em diferentes unidades de medida.

Suponha-se que se pretendia agrupar n indivíduos (poços). Os algoritmos de agrupamento operam, geralmente, sobre dois tipos de estrutura de dados: o primeiro tipo apresenta os indivíduos sob a forma de uma matriz de dimensão $n \times p$ correspondendo as n linhas aos poços e as p colunas aos seus atributos ou características; o segundo tipo consiste numa apresentação sob a forma de um quadro de dimensão $n \times n$ cujos elementos medem as proximidades entre cada par de poços. Estas proximidades poderão ser semelhanças (quando medem o grau de semelhança entre cada par de poços) ou distâncias (quando medem o grau de afastamento ou diferença).

Quando os dados se encontram na primeira forma, há que atender ao tipo de variáveis (contínuas, ordinais, nominais ou binárias) para se escolher o algoritmo de agrupamento adequado. Quando, adicionalmente, as variáveis se apresentam definidas em diferentes escalas de medida e se aplica a análise de agrupamentos sem uma padronização prévia, qualquer medida de semelhança/distância vai refletir, sobretudo o peso das variáveis que apresentam maiores valores e maior dispersão.

Para anular este efeito, o processo mais utilizado consiste na padronização das variáveis, um processo relativamente simples uma vez conhecidas as médias e os desvios-padrão das variáveis e que consiste na sua transformação em novas variáveis ($Z = (X - m)/s$) com média nula e desvio-padrão unitário.

2.2.2 Técnicas de Análise de Conglomerados

Na aplicação do método, é necessário identificar a técnica de análise mais apropriada. É possível dividir as técnicas disponíveis em vários grupos:

1. Técnicas de otimização: é definido um critério de agrupamento e a sua otimização indica qual deverá ser o grupo onde cada caso será incluído, pressupondo que

todos os casos pertencem a um número k predeterminado de grupos;

2. Técnicas hierárquicas: que se podem subdividir em técnicas aglomerativas e divisivas, ambas partindo de uma matriz de semelhanças ou dissimilaridades (distâncias) entre os casos. Estes métodos conduzem a uma hierarquia de partições $P_1, P_2, ..., P_n$ do conjunto de n objetos em $1, 2, ..., n$ grupos. Os métodos dizem-se hierárquicos porque, para cada par de partições, P_i e P_{i+1}, cada grupo da partição P_{i+1} está incluído num grupo da partição P_i.

3. Técnicas de densidade (*density or mode-seeking*): os grupos são formados através da procura de regiões que contenham uma concentração relativamente densa de casos.

4. Outras técnicas: incluem aquelas em que se permite a s obreposição dos grupos (*fuzzy clusters*).

2.2.3 Distâncias e Coeficientes de Similaridade para Pares de Indivíduos

Um conjunto de regras, conhecidas como medidas de similaridade, é utilizado no agrupamento dos indívíduos. Assume-se que a distância entre os pontos (objetos ou variáveis) reflete a similaridade (ou dissimilaridade) de suas propriedades. Portanto, quanto mais próximos estiverem os pontos no espaço amostral, mais similares (ou, respectivamente, dissimilares) eles serão. Os resultados são fornecidos na forma de dendogramas, que agrupam objetos ou variáveis em função das similaridades (ou, respectivamente, dissimilaridades) [3, 21]. Dentre as principais medidas de similaridade, podemos citar: Distância Euclidiana, Quadrado da Distância Euclidiana, Distância Absoluta ou *City-Block Metric* e Distância de Minkowski. Neste trabalho, os indivíduos são os poços e as variáveis são os perfis utilizados na avaliação das formações geológicas.

A escolha de uma medida de similaridade envolve, muitas vezes, uma grande subjetividade. Importantes considerações envolvem, por exemplo, a natureza das variáveis (se estas são discretas, contínuas ou binárias) e as escalas de medida (se as variáveis são nominais, ordinais).

Sejam $\boldsymbol{x}$ e $\boldsymbol{y}$ dois pontos arbitrários, com coordenadas $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$ e $\boldsymbol{y} = (y_1, y_2, ..., y_p)'$.

A distância euclidiana entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada por

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}. \tag{2.1}$$

O quadrado da distância euclidiana entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dado por

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p}(x_i - y_i)^2. \tag{2.2}$$

A distância absoluta ou *City-Block Metric* entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada por

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \mid x_i - y_i \mid . \tag{2.3}$$

A generalização da distância euclidiana e da distância absoluta entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada pela distância de Minkowski

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{p} \mid x_i - y_i \mid^m \right)^{1/m}, \text{ com } m > 0. \tag{2.4}$$

A distância de Mahalanobis entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada por

$$d(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y}), \text{ onde } \boldsymbol{\Sigma} \text{ é a matriz de covariâncias.} \tag{2.5}$$

A distância de Chebishev entre $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada por

$$d(\boldsymbol{x}, \boldsymbol{y}) = \max |(x_i - y_i)|. \tag{2.6}$$

Apesar da sua importância, a distância euclidiana, bem como as outras medidas de distância, apresenta vários problemas de utilização, tornando mais importante o efeito que as diferenças de escala das variáveis provocam sobre o valor das distâncias. As variáveis que apresentam variações e unidades de medida elevadas, facilmente anularão o efeito das outras variáveis. Para resolver este problema é comum a prática de padronização das variáveis, de modo a tornar a sua média nula e o seu desvio-padrão unitário.

A Análise de Conglomerados facilita o estudo de grandes conjuntos de dados, tornando-se, então, em particular, uma boa ferramenta no estudo de dados relativos a reservatórios.

2.2.4 Aplicação

```
# Descrição:
# Dados referentes às variáveis DT, GR, ILD e RHOB dos poços
# 1RJS0019RJ, 3NA001ARJS, 3NA0002RJS, 3NA0003RJS, 3NA0004RJS e
# 3NA021BRJS.

# Questão de interesse:
# Verificar se os poços podem ser agrupados ou classificados em
# conglomerados informativos, com base nos valores das
# variáveis DT, GR, ILD e RHOB.

# Leitura dos dados:
dados<-read.table("conglomerados.txt", header=TRUE)

# Excluindo a primeira coluna (nomes dos poços):
caracteristicas<-as.matrix(dados[,4:7])
caracteristicas
       DT      GR     ILD   RHOB
1 70.0430 42.4961  1.9849 2.4941
2 87.7578 61.1250  9.2383 2.3242
3 84.6250 48.9492 14.7455 2.4353
4 80.3008 83.4170  2.0567 2.3102
5 93.3633 60.0556 30.2517 2.2461
6 72.5547 48.1361  3.2778 2.5721

# Calculando o desvio padrão de cada variável:
caracteristicas.dp<-sqrt(apply(caracteristicas,2,var))
caracteristicas.dp
       DT         GR        ILD       RHOB
8.9706393 14.6776434 11.0032193  0.1242258

# Padronização dos dados (dividindo cada valor pelo desvio padrão
# da respectiva coluna):
caracteristicas.std<-sweep(caracteristicas,2,caracteristicas.dp,FUN="/")
caracteristicas.std
         DT       GR        ILD     RHOB
1  7.808028 2.895294 0.1803927 20.07716
2  9.782781 4.164497 0.8395997 18.70948
3  9.433553 3.334950 1.3401078 19.60382
```

```
4  8.951514 5.683269 0.1869180 18.59679
5 10.407653 4.091638 2.7493499 18.08079
6  8.088019 3.279552 0.2978946 20.70505

# Calculando a matriz de distâncias (usando distância euclidiana):
caracteristicas.dist<-dist(caracteristicas.std)
caracteristicas.dist
          1         2         3         4         5
2 2.7956402
3 2.0987165 1.3639870
4 3.3573649 1.8537447 2.8444575
5 4.3329325 2.1066977 2.4139268 3.3890916
6 0.7963043 2.8162122 2.0279118 3.3136909 4.3516122

# Realizando o agrupamento através do método do complete linkage:
clust.complete<-hclust(caracteristicas.dist, method="complete")

# Fazendo o dendograma:
win.graph()
plclust(clust.complete, xlab="", ylab="", main="Dendograma do agrupamento
dos poços", sub="Complete Linkage", labels=dimnames(caracteristicas)[[1]])
```

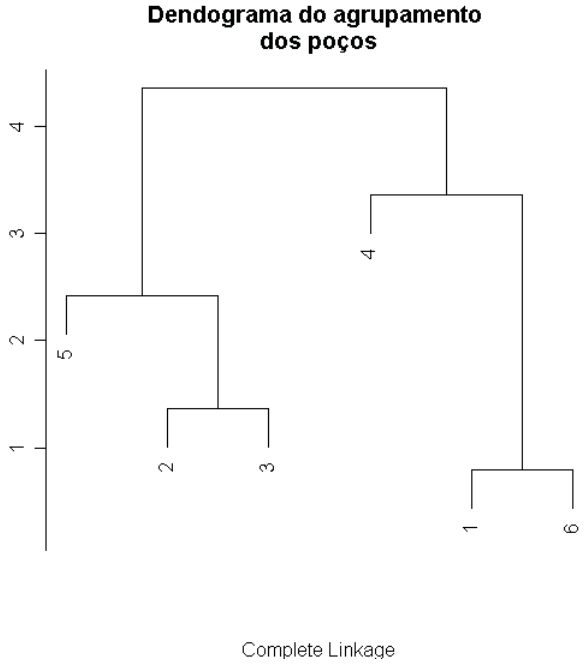


Figura 2.1: Dendograma do agrupamento dos 6 poços.

```
# Solução para quatro grupos usando a função "cutree" (criação
# de um vetor contendo o conglomerado de cada poço):
caracter.4grupos<-cutree(clust.complete, k=4)
caracter.4grupos
1 2 3 4 5 6
1 2 2 3 4 1

# Identificando os membros de cada conglomerado:
for(i in 1:4)print(dimnames(caracteristicas)[[1]][caracter.4grupos==i])
[1] "1" "6"
[1] "2" "3"
[1] "4"
[1] "5"

# Cálculo dos coeficientes de fusão:
coef.fusao.complete<-clust.complete$height
coef.fusao.complete
[1] 0.7963043 1.3639870 2.4139268 3.3573649 4.3516122
```

2.3 Análise de Componentes Principais e Análise Fatorial

Quando dados multivariados são coletados, é comum se encontrar algumas variáveis correlacionadas entre si. Uma implicação dessas correlações é a existência de redundância nas informações contidas nos dados. A Análise de Componentes Principais é uma técnica através da qual são obtidas combinações lineares das variáveis correlacionadas [19]. Ela consiste de uma técnica que tem vasta aplicação, particularmente em Geologia, e está diretamente relacionada com a transformação de variáveis, através do cálculo dos autovalores e correspondentes autovetores da matriz de variâncias-covariâncias ou da matriz de correlação entre variáveis, de forma a preservar a variabilidade total.

A primeira etapa na análise por componentes principais é considerar a matriz de variâncias-covariâncias ($\mathbf{\Sigma}$). Quando as variáveis, devido a escalas diferentes de mensurações empregadas, não podem ser diretamente comparadas, torna-se necessário, preliminarmente, fazer uma padronização, de modo que as variáveis transformadas passam

a ter média zero e variância unitária. Nesses casos, com variáveis padronizadas, a matriz de variâncias-covariâncias e a de correlação tornam-se idênticas. Como tal padronização acarreta uma forte influência na estrutura da matriz de variâncias-covariâncias e, conseqüentemente, nos resultados da análise, a sua utilização deve ser criteriosa levando sempre em conta a natureza dos dados geológicos em estudo e o enfoque que se pretende dar.

A etapa seguinte deve ser o cálculo dos autovalores e correspondentes autovetores dessa matriz. O primeiro autovalor a ser determinado corresponderá à maior porcentagem da variabilidade total presente e assim sucessivamente. Geralmente os dois ou três primeiros autovalores encontrados explicarão a maior parte da variabilidade presente.

Os autovetores correspondem às componentes principais e são o resultado da carga das variáveis originais em cada um deles. Tais cargas podem ser considerados como uma medida da relativa importância de cada variável em relação às componentes principais e os respectivos sinais, se positivos ou negativos, indicam relações diretamente e inversamente proporcionais.

A matriz de cargas de cada variável nas componentes principais ao ser multiplicada pela matriz original de dados fornecerá a matriz de escores de cada caso em relação às componentes principais.

Algebricamente, componentes principais são combinações lineares de p variáveis aleatórias $X_1, X_2, ..., X_p$. Geometricamente, essas combinações lineares representam a seleção de um novo sistema de coordenadas obtido pela rotação do sistema original com coordenadas $X_1, X_2, ..., X_p$. Os novos eixos representam as direções com variabilidade máxima e proporcionam uma descrição mais simples da estrutura de covariâncias.

Assim, as componentes principais dependem apenas da matriz de covariâncias $\mathbf{\Sigma}$ (ou da matriz de correlações $\boldsymbol{\rho}$) de $X_1, X_2, ..., X_p$.

Consideremos o vetor aleatório $\boldsymbol{X}' = (X_1, X_2, ..., X_p)$, com matriz de covariâncias $\mathbf{\Sigma}$, positiva definida, e autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Sejam as combinações lineares

$$\begin{array}{rcrcl} Y_1 & = & \boldsymbol{a}_1^{'}\boldsymbol{X} & = & a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 & = & \boldsymbol{a}_2^{'}\boldsymbol{X} & = & a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \vdots & & & & \vdots \\ Y_p & = & \boldsymbol{a}_p^{'}\boldsymbol{X} & = & a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p. \end{array} \tag{2.7}$$

Usando o fato de que as combinações lineares $\boldsymbol{Z} = \boldsymbol{C}\boldsymbol{X}$, onde $\boldsymbol{C}$ é uma matriz $k \times p$, têm

$$\boldsymbol{\mu_z} = E(\boldsymbol{Z}) = E(\boldsymbol{C}\boldsymbol{X}) = C\boldsymbol{\mu_x}$$

e

$$\boldsymbol{\Sigma_z} = Cov(\boldsymbol{Z}) = Cov(\boldsymbol{C}\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{\Sigma_x}\boldsymbol{C}^{'},$$

obtemos que

$$Var(Y_i) = \boldsymbol{a}_i^{'}\boldsymbol{\Sigma}\boldsymbol{a}_i \quad i = 1, 2, ..., p \tag{2.8}$$

$$Cov(Y_i, Y_k) = \boldsymbol{a}_i^{'}\boldsymbol{\Sigma}\boldsymbol{a}_k \quad i, k = 1, 2, ..., p \ . \tag{2.9}$$

As componentes principais são as combinações lineares não correlacionadas $Y_1, Y_2, ..., Y_p$ cujas variâncias em (2.8) sejam as maiores possíveis.

A primeira componente principal é a combinação linear com variância máxima, ou seja, que maximiza $Var(Y_1) = \boldsymbol{a}_1^{'}\boldsymbol{\Sigma}\boldsymbol{a}_1$. Note que $Var(Y_1) = \boldsymbol{a}_1^{'}\boldsymbol{\Sigma}\boldsymbol{a}_1$ pode sofrer um acréscimo se $\boldsymbol{a}_1$ for multiplicada por alguma constante. Para eliminar essa indeterminação, é conveniente restringir atenção a vetores unitários.

Definimos, então

- Primeira componente principal: combinação linear $\boldsymbol{a}_1^{'}\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_1^{'}\boldsymbol{X})$ sujeito a $\boldsymbol{a}_1^{'}\boldsymbol{a}_1 = 1$;

- Segunda componente principal: combinação linear $\boldsymbol{a}_2^{'}\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_2^{'}\boldsymbol{X})$ sujeito a $\boldsymbol{a}_2^{'}\boldsymbol{a}_2 = 1$ e $Cov(\boldsymbol{a}_1^{'}\boldsymbol{X}, \boldsymbol{a}_2^{'}\boldsymbol{X}) = 0$;

- i-ésima componente principal: combinação linear $\boldsymbol{a}_i^{'}\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_i^{'}\boldsymbol{X})$ sujeito a $\boldsymbol{a}_i^{'}\boldsymbol{a}_i = 1$ e $Cov(\boldsymbol{a}_i^{'}\boldsymbol{X}, \boldsymbol{a}_k^{'}\boldsymbol{X}) = 0$, para $k < i$.

Sendo $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ os pares de autovalores e autovetores normalizados da matriz de covariâncias $\boldsymbol{\Sigma}$, associada ao vetor aleatório $\boldsymbol{X}' = (X_1, X_2, ..., X_p)$, onde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, temos que a i-ésima componente principal é dada por

$$Y_i = \boldsymbol{e}_i' \boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, ..., p. \tag{2.10}$$

Como $\boldsymbol{\Sigma}$ é uma matriz positiva definida e associada aos autovalores normalizados $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_p$, temos que

$$\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a}}{\boldsymbol{a}' \boldsymbol{a}} = \lambda_1, \text{ quando } \boldsymbol{a} = \boldsymbol{e}_1.$$

Mas $\boldsymbol{e}_1' \boldsymbol{e}_1 = 1$.

Logo, temos que

$$\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a}}{\boldsymbol{a}' \boldsymbol{a}} = \lambda_1 = \frac{\boldsymbol{e}_1' \boldsymbol{\Sigma} \boldsymbol{e}_1}{\boldsymbol{e}_1' \boldsymbol{e}_1} = \boldsymbol{e}_1' \boldsymbol{\Sigma} \boldsymbol{e}_1 = Var(Y_1).$$

Além disso, para $\boldsymbol{a} = \boldsymbol{e}_{k+1}$, com $\boldsymbol{e}_{k+1}' \boldsymbol{e}_i = 0$, $i = 1, 2, ..., k$ e $k = 1, 2, ..., p-1$, temos:

$$\max_{\boldsymbol{a} \perp \boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k} \frac{\boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a}}{\boldsymbol{a}' \boldsymbol{a}} = \lambda_{k+1}$$

e

$$\frac{\boldsymbol{e}_{k+1}' \boldsymbol{\Sigma} \boldsymbol{e}_{k+1}}{\boldsymbol{e}_{k+1}' \boldsymbol{e}_{k+1}} = \boldsymbol{e}_{k+1}' \boldsymbol{\Sigma} \boldsymbol{e}_{k+1} = Var(Y_{k+1}).$$

Mas $\boldsymbol{e}_{k+1}' (\boldsymbol{\Sigma} \boldsymbol{e}_{k+1}) = \lambda_{k+1} \boldsymbol{e}_{k+1}' \boldsymbol{e}_{k+1} = \lambda_{k+1}$. Portanto, $Var(Y_{k+1}) = \lambda_{k+1}$. Resta mostrar que se $\boldsymbol{e}_i$ é perpendicular a $\boldsymbol{e}_k$, temos que $Cov(Y_i, Y_k) = 0$. Note que os autovetores de $\boldsymbol{\Sigma}$ são ortogonais se todos os autovalores $\lambda_1, \lambda_2, ..., \lambda_p$ forem distintos. Caso contrário, os autovetores correspondentes aos autovalores iguais devem ser escolhidos de maneira que sejam ortogonais. Então, para quaisquer dois autovetores $\boldsymbol{e}_i$ e $\boldsymbol{e}_k$, com $i \neq k$, temos que $\boldsymbol{e}_i' \boldsymbol{e}_k = 0$. Se $\boldsymbol{\Sigma} \boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k$, multiplicando ambos os membros por $\boldsymbol{e}_i'$, temos que

$$Cov(Y_i, Y_k) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_k = \boldsymbol{e}_i' \lambda_k \boldsymbol{e}_k = \lambda_k \boldsymbol{e}_i' \boldsymbol{e}_k = 0,$$

para qualquer $i \neq k$.

Desta forma, temos

$$\begin{aligned} &Var(Y_i) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i, \text{ com } i = 1, 2, ..., p, \\ &Cov(Y_i, Y_k) = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_k = 0, \text{ com } i \neq k. \end{aligned} \tag{2.11}$$

Devemos salientar que não temos unicidade na escolha de cada autovetor e_i, e consequentemente Y_i não é única, mesmo quando todos os autovalores são distintos.

Nada é tratado a respeito de probabilidade, testes de hipóteses, etc, porque Análise de Componentes Principais é um procedimento descritivo. Entretanto, assume algumas das características de procedimentos estatísticos, quando são tomadas decisões de descartar algumas das novas variáveis ou componentes. Análise de Componentes Principais pertence a uma categoria de técnicas, incluindo Análise de Conglomerados, posteriormente citada, em que sua utilidade é julgada pela performance e não pelas considerações teóricas.

Análise Fatorial é um dos procedimentos multivariados mais amplamente utilizados, e pode ser considerado uma extensão de análise de Componentes Principais. Ambas podem ser vistas como tentativas de aproximar a matriz de variâncias-covariâncias amostral ($\hat{\mathbf{\Sigma}}$), por uma matriz mais simples, obtida por meio da estrutura de autovalores e autovetores. Entretanto, a aproximação baseada em modelos de análise fatorial é mais elaborada [12]. Por isso, ela é comumente considerada como uma técnica estatística, no sentido de admitir um modelo. A análise fatorial considera um conjunto de suposições a respeito da natureza da população de onde as amostras foram extraídas. Estas suposições melhoram o raciocínio para a execução das operações e a maneira com que os resultados são interpretados.

Convém enfatizar que, apesar de possuírem muitos aspectos em comum, a análise das componentes principais não é "sinônimo" de análise fatorial ou análise dos fatores e essa confusão terminológica deve ser evitada. A primeira análise consiste numa transformação linear de p variáveis originais em p novas variáveis, onde cada nova variável é uma combinação linear das variáveis antigas [7]. Isto ocorre de tal modo que a primeira nova variável computada seja responsável pela maior variação possível existente no conjunto de dados, a segunda pela maior variação possível restante e assim por diante, até que toda a variação do conjunto tenha sido explicada. Na análise fatorial, supõe-se que as relações existentes dentro de um conjunto de p variáveis seja o reflexo das correlações de cada uma dessas variáveis com m fatores, mutuamente não correlacionados entre si. A suposição usual é que $m < p$.

A análise fatorial não se refere a uma única técnica estatística, mas a uma variedade de técnicas relacionadas para tornar os dados observados mais facilmente (e

diretamente) interpretados. Isto é feito analisando-se os inter-relacionamentos entre as variáveis de tal modo que estas possam ser descritas, convenientemente, por um grupo de categorias básicas, em número menor que as variáveis originais, chamados fatores. Assim, o objetivo da análise fatorial é a parcimônia, procurando definir o relacionamento entre as variáveis de modo simples e usando um número de fatores menor do que o número original de variáveis.

Mais precisamente, um fator é um construto, uma entidade hipotética, uma variável não observada, que se supõe estar subjacente a testes, escalas, ítens e, de fato, medidas de qualquer espécie. Como construtos, os fatores apenas possuem realidade no fato de explicarem a variância de variáveis observadas, tal como se revelam pelas correlações entre as variáveis sendo analisadas, ou seja, a única realidade científica que os fatores possuem vem das correlações entre as variáveis que estão sendo pesquisadas [23]. Se os resultados de indivíduos em ítens ou testes caminham juntos, então, na medida em que existam correlações substanciais entre eles, está definido um fator.

2.3.1 Modelo Fatorial Ortogonal

Observemos o vetor aleatório $\boldsymbol{X}$, com p componentes, com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. O modelo fatorial postula que $\boldsymbol{X}$ é linearmente dependente sobre variáveis aleatórias não observáveis $F_1, F_2, ..., F_m$, chamadas fatores comuns, e p fontes adicionais de variação $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$, chamados de erros ou, às vezes, fatores específicos. Em particular, o modelo de análise fatorial é:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ \vdots \quad &\quad \vdots \qquad\qquad\qquad \vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

Ou, em notação matricial,

$$(\boldsymbol{X} - \boldsymbol{\mu})_{(p\times 1)} = \boldsymbol{L}_{(p\times m)}\boldsymbol{F}_{(m\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)}, \tag{2.12}$$

onde

$$\begin{aligned}\mu_i &= \text{média da variável } X_i;\\ \varepsilon_i &= i\text{-ésimo fator específico;}\\ F_j &= j\text{-ésimo fator comum;}\\ l_{ij} &= \text{carga da } i\text{-ésima variável sobre o } j\text{-ésimo fator.}\end{aligned}$$

A matriz $\boldsymbol{L}$ é a matriz de cargas. Note que o i-ésimo fator específico é associado apenas com a i-ésima resposta X_i. Os p desvios $X_1-\mu_1,\ldots,X_p-\mu_p$ são expressos em termos de $p+m$ variáveis aleatórias $F_1,F_2,\ldots,F_m,\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_p$ que não são observadas.

Com tantas quantidades não observadas, uma verificação direta do modelo fatorial das observações em $X_1,X_2,\ldots,X_p$ é impossível. Entretanto, com algumas suposições sobre os vetores aleatórios $\boldsymbol{F}$ e $\boldsymbol{\varepsilon}$, o modelo em (2.12) implica certas estruturas de covariância, que podem ser checadas.

Neste trabalho, assumimos que:

$i) E(\boldsymbol{F})=\boldsymbol{0}_{(m\times 1)}, Cov(\boldsymbol{F})=E(\boldsymbol{F}\boldsymbol{F}')=\boldsymbol{I}_{(m\times m)}$

$$ii)\quad E(\boldsymbol{\varepsilon})=\boldsymbol{0}_{(p\times 1)},\quad Cov(\boldsymbol{\varepsilon})=E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')=\boldsymbol{\Psi}_{(p\times p)}=\begin{bmatrix}\psi_1 & 0 & \cdots & 0\\ 0 & \psi_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \psi_p\end{bmatrix} \tag{2.13}$$

e que $\boldsymbol{F}$ e $\boldsymbol{\varepsilon}$ são não correlacionados. Logo,

$$Cov(\boldsymbol{\varepsilon},\boldsymbol{F})=E(\boldsymbol{\varepsilon}\boldsymbol{F}')=\boldsymbol{0}_{(p\times m)}.$$

Essas suposições e a relação (2.12) constituem o modelo fatorial ortogonal [12].

O modelo fatorial ortogonal implica uma estrutura de covaiância para $\boldsymbol{X}$. De (2.12) temos que

$$\begin{aligned}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})' &= (\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\varepsilon})(\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\varepsilon})'\\ &= (\boldsymbol{L}\boldsymbol{F}+\boldsymbol{\varepsilon})(\boldsymbol{F}'\boldsymbol{L}'+\boldsymbol{\varepsilon}')\\ &= \boldsymbol{L}\boldsymbol{F}\boldsymbol{F}'\boldsymbol{L}'+\boldsymbol{\varepsilon}\boldsymbol{F}'\boldsymbol{L}'+\boldsymbol{L}\boldsymbol{F}\boldsymbol{\varepsilon}'+\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\end{aligned}$$

e então

$$\begin{aligned}\boldsymbol{\Sigma} &= Cov(\boldsymbol{X})=E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\\ &= \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{F}')\boldsymbol{L}'+E(\boldsymbol{\varepsilon}\boldsymbol{F}')\boldsymbol{L}'+\boldsymbol{L}E(\boldsymbol{F}\boldsymbol{\varepsilon}')+E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\ ,\\ &= \boldsymbol{L}\boldsymbol{L}'+\boldsymbol{\Psi},\end{aligned}$$

de acordo com (2.13).

Além disso, de (2.12), temos que $Cov(\boldsymbol{X}, \boldsymbol{F}) = E(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{F}' = \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{F}') + E(\boldsymbol{\varepsilon}\boldsymbol{F}') = \boldsymbol{L}$.

Uma matriz de cargas fatoriais é um dos produtos finais da análise fatorial. Uma carga fatorial é um coeficiente - um valor, positivo ou negativo, geralmente menor do que 1 - que expressa o quanto uma variável observada está carregada ou saturada em um fator. Em outras palavras, quanto maior for a carga em cima de um fator, mais a variável se identifica com o fator.

Em síntese, a análise fatorial é essencialmente um método para determinar o número de fatores existentes em um conjunto de dados, como também para identificar quais variáveis pertencem a quais fatores, e em que extensão as variáveis estão saturadas com o que quer que seja o fator [23].

As cargas fatoriais obtidas são, com efeito, reduções de dados muito mais complexos a tamanho manuseável para que o pesquisador possa interpretar melhor os resultados.

2.3.2 Aplicação

Podemos, por exemplo, fazer uso da Análise de Componentes Principais para a análise e interpretação dos conglomerados obtidos no ítem 2.2.4:

```
caracter.cp<-prcomp(caracteristicas.std)
caracter.cp
Standard deviations:
[1] 1.6585129 1.0480256 0.2762087 0.2732877
Rotation:
            PC1        PC2        PC3        PC4
DT    0.5751691 -0.1970391  0.4355326 -0.6638279
GR    0.3481305  0.7659868 -0.4829347 -0.2425768
ILD   0.4668620 -0.5774756 -0.6533593  0.1472530
RHOB -0.5744785 -0.2023903 -0.3875656 -0.6919578

x<-caracter.cp$x[,1]
x
          1           2           3           4           5           6
 -1.88352133  0.78760027  0.01783371  0.59824229  2.37440379 -1.89455871
```

```
y<-caracter.cp$x[,2]
y
          1           2           3           4           5           6
 -0.2494655  0.2297504 -0.8068970  1.9566189 -0.9247754 -0.2052314
#-----------------------------------------------------------------------#
win.graph() plot(x,y,xlab="Primeira Componente Principal",
ylab="Segunda Componente Principal", main="Usando CP para a
vizualização dos conglomerados",type="n")
text(x,y,labels=as.character(caracter.4grupos))
```

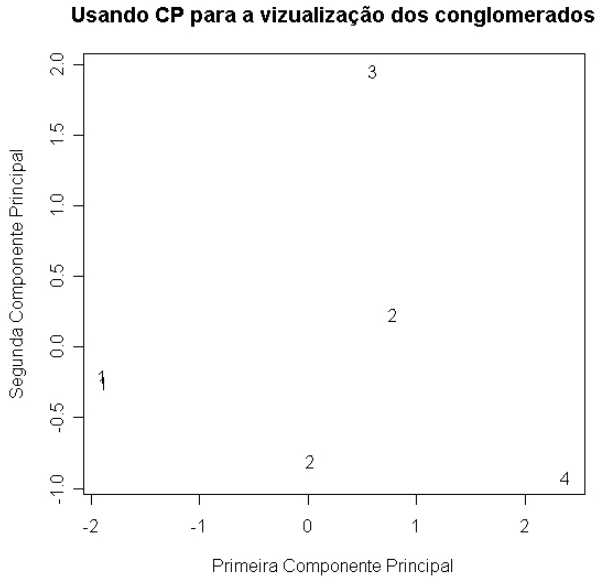


Figura 2.2: Disposição dos conglomerados, segundo as duas primeiras componentes principais.

2.4 Análise Discriminante

Um dos procedimentos multivariados mais amplamente utilizados na ciência da terra é a função discriminante [7].

Quando se dispõe de algumas informações a respeito de determinados indivíduos, obtidas através de um certo número de variáveis, é razoável questionar se estas variáveis podem ser utilizadas para definir alguma relação entre os membros de um grupo. O objetivo da Análise Discriminante é fazer uma combinação linear destas variáveis, de tal forma que as diferenças entre grupos pré-definidos sejam maximizadas.

Dois processos são de grande relevância no estudo de Análise Discriminante: a discriminação e a classificação. Por este motivo, dedicaremos o próximo tópico a esses processos.

Discriminação e classificação são técnicas multivariadas que objetivam a separação de conjuntos de objetos distintos (ou observações) e que alocam novos objetos (observações) em grupos previamente definidos. A Análise Discriminante é especialmente exploratória por natureza. Processos de Classificação são menos exploratórios no sentido de que eles conduzem a regras bem definidas, que podem ser usadas para associar novos objetos. Classificação ordinariamente exige mais estrutura de problema do que a discriminação requer.

2.4.1 Separação e Classificação no Caso de Duas Populações

Os objetivos imediatos da discriminação e classificação, respectivamente, são:

Objetivo 1: Descrever, algebricamente ou graficamente, as características diferenciais dos objetos (observações) de muitas coleções (populações) conhecidas.

Objetivo 2: Classificar objetos (observações) dentro de duas classes designadas.

Nós devemos seguir a rotina e usar o termo *discriminação* para nos referirmos ao objetivo 1. Esta terminologia foi introduzida por R. A. Fisher no primeiro tratamento moderno de problemas separativos [9]. Um termo mais descritivo para este objetivo, entretanto, é *separação*. Nós devemos nos referir ao segundo objetivo como *classificação* ou *alocação*.

A função que separa os objetos pode algumas vezes servir como alocadora, e, reciprocamente, a regra que aloca objetos pode sugerir um processo discriminatório. Na prática, os objetivos 1 e 2 freqüentemente são sobrepostos, e a distinção entre separação e alocação chega a ser obscura.

O objetivo da Análise Discriminante é combinar as variáveis consideradas de forma que seja obtida uma nova e única variável composta, denominada fator discriminante.

No final do processo, espera-se que cada grupo tenha distribuição normal de fatores discriminantes. Desta forma, o grau de coincidência entre as distribuições do

fator discriminante pode ser usada como uma medida do sucesso da técnica. Quanto menos coincidência houver, mais eficiente será a separação dos grupos. Observemos, por exemplo, a figura 2.3:

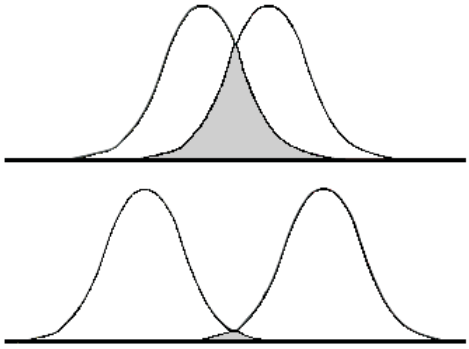


Figura 2.3: Grau de coincidência entre as distribuições do fator discriminante.

Os fatores discriminantes são calculados a partir da *função discriminante*, dada por:

$$D_k = w_1 Z_{1k} + w_2 Z_{2k} + ... + w_i Z_{ik} + ... + w_p Z_{p_k}, \quad k = 1, ..., n$$

onde

- D_k é o fator discriminante;
- w_i é a carga (coeficiente) para a variável i;
- Z_i é o fator padronizado para a variável i (com média igual a zero e desvio padrão igual a um).

Desta forma, um fator discriminante é uma combinação linear dos pesos das variáveis discriminantes.

A padronização das variáveis assegura a eliminação da diferença de escalas entre as variáveis.

Quando os pesos absolutos das variáveis são padronizados, estas podem ser ordenadas em termos do seu poder de discriminação, sendo o maior peso associado à variável discriminante mais poderosa. Variáveis com pesos altos são aquelas que mais contribuem na diferenciação dos grupos.

Um procedimento de classificação pode resultar em algumas classificações errôneas. Assim, para que um procedimento de classificação seja considerado bom, as

chances, ou probabilidades, de ocorrerem estas classificações errôneas devem ser pequenas. Além disso, existem outras características que uma regra de classificação "ótima" deve possuir.

Pode acontecer de uma classe ou população ter uma chance muito maior de ocorrência do que outra, pelo fato de uma das duas populações ser relativamente maior que outra. Uma regra ótima de classificação deve levar em consideração essas probabilidades de ocorrência *a priori*.

É conveniente rotular as classes por π_1 e π_2. Se, por exemplo, classificar um objeto da classe π_1 como sendo da classe π_2 representar um erro mais sério que classificar um objeto da classe π_2 como sendo da classe π_1, então deve-se ter cautela no intuito de evitar a ocorrência do primeiro erro. Outro aspecto que deve ser observado é o custo. Um bom procedimento deve, sempre que possível, considerar os custos associados às classificações errôneas.

Sejam $f_1(\boldsymbol{x})$ e $f_2(\boldsymbol{x})$ as funções densidade de probabilidade associadas com o vetor aleatório da variável $\boldsymbol{X}$ de dimensão $p \times 1$ para as populações π_1 e π_2, respectivamente. A observação $\boldsymbol{x}$ deve ser associada a uma das duas populações, π_1 ou π_2. Seja Ω o espaço amostral, isto é, a coleção de todas as possíveis observações $\boldsymbol{x}$. Seja R_1 o conjunto dos valores $\boldsymbol{x}$ para os quais os objetos são classificados como pertencentes a π_1 e $R_2 = \Omega - R_1$ os valores remanescentes de $\boldsymbol{x}$ para os quais os objetos são classificados como pertencentes a π_2. Como cada objeto deve ser associado a apenas uma das duas populações, a intersecção entre os conjuntos R_1 e R_2 é vazia. Além disso, temos que $R_1 \bigcup R_2 = \Omega$.

A probabilidade condicional, $P(2|1)$, de classificar uma objeto como sendo de π_2 quando ele é de π_1 é dada por

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(\boldsymbol{x})d\boldsymbol{x}. \tag{2.14}$$

Da mesma maneira, temos que a probabilidade condicional, $P(1|2)$, de classificar uma objeto como sendo de π_1 quando ele é de π_2 é dada por

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}. \tag{2.15}$$

Seja p_1 a probabilidade a priori de π_1 e p_2 a probabilidade a priori de π_2, onde $p_1 + p_2 = 1$. Assim, as probabilidades totais de classificar objetos corretamente ou

incorretamente pode ser obtida a partir do produto das probabilidades de classificação condicional pelas probabilidades de classificação *a priori*:

- A probabilidade P(observação ser classificada corretamente como sendo de π_1) é igual a P(observação ser de π_1 e ser classificada como sendo de π_1), e é dada por:

$$P(\boldsymbol{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1;$$

- A probabilidade P(observação ser classificada incorretamente como sendo de π_1) é igual a P(observação ser de π_2 e ser classificada como sendo de π_1), e é dada por:

$$P(\boldsymbol{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2;$$

- A probabilidade P(observação ser classificada corretamente como sendo de π_2) é igual a P(observação ser de π_2 e ser classificada como sendo de π_2), e é dada por:

$$P(\boldsymbol{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2;$$

- A probabilidade P(observação ser classificada incorretamente como sendo de π_2) é igual a P(observação ser de π_1 e ser classificada como sendo de π_2), e é dada por:

$$P(\boldsymbol{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1.$$

Os custos de classificação errônea pode ser definido pela matriz de custos:

Tabela 2.1: Custos de classificação errônea.

	Classificado como	
População Real	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

Os custos são: zero para classificação correta, $c(1|2)$ quando uma observação de π_2 é classificada incorretamente como sendo de π_1, e $c(2|1)$ quando uma observação de π_1 é classificada incorretamente como sendo de π_2.

Para qualquer regra, o *custo esperado por classificação incorreta* (ECM) é obtido através da multiplicação de $c(2|1)$ e $c(1|2)$ pelas suas respectivas probabilidades de ocorrência, obtidas anteriormente. Então, temos:

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2. \tag{2.16}$$

Uma regra de classificação eficiente deve ter um ECM tão pequeno quanto seja possível.

2.4.2 Aplicação

```
# Questão de interesse:

# Classificar cada observação como referente ao poço 1RJS0019RJ ou
# ao poço 3NA0002RJS, através da aplicação da Análise Discriminante
# aos dados das variáveis DT e ILD dos mesmos.

# Entrada de dados

# No exemplo a seguir, as primeiras 30 linhas contêm informações #
referentes à amostra do poço 1RJS0019RJ, e as demais linhas #
contêm informações referentes à amostra do poço 3NA0002RJS.

mdados<-read.table("po1e2.txt",header=TRUE)
mdados<-cbind(mdados$DT,mdados$ILD)
mdados
          [,1]     [,2]
 [1,]  84.1809   2.0278
 [2,]  86.7461   1.6348
 [3,]  89.6758   1.6054
 [4,]  89.6680   2.1108
 [5,]  85.0625   2.2480
 [6,]  92.8789   2.7148
 [7,]  84.7344   2.7734
 [8,]  94.3552   2.9285
 [9,]  89.8047   4.8696
[10,] 110.2070   9.6970
[11,]  93.5312   5.3892
[12,]  77.9180   2.7256
[13,]  86.4336   1.7183
[14,]  84.6033   1.7896
[15,]  91.7109   4.8125
[16,]  94.6211   4.9150
[17,]  76.0000   3.6597
[18,]  84.1450   2.0310
[19,]  86.8750   1.7886
```

```
[20,]  97.2070   6.4727
[21,]  83.5627   1.9921
[22,]  79.5117   2.1458
[23,]  79.3516   3.0627
[24,]  74.8906   4.2173
[25,]  79.6875  21.6758
[26,]  84.9375  32.1875
[27,]  98.4651  13.2656
[28,]  80.3125  33.9375
[29,]  78.5195   1.8826
[30,]  83.2959   2.0723
[31,]  91.2695   1.8425
[32,]  90.9714   1.4395
[33,]  95.6133   1.6209
[34,]  85.5479   1.9060
[35,]  90.6836   1.6438
[36,]  92.8125   1.4741
[37,]  95.5234   1.8086
[38,] 102.0547   2.9922
[39,]  96.8945   4.3594
[40,] 100.7148   2.5627
[41,]  93.0625   4.2539
[42,]  65.7500  18.1797
[43,]  86.2109   8.6641
[44,] 104.4895 274.0000
[45,]  86.3750  39.0611
[46,]  76.0117  44.6250
[47,]  96.9492 353.0000
[48,]  78.0625  10.8945
[49,]  87.8750   3.8750
[50,]  71.0391   7.2578
[51,]  99.6719   3.3254
[52,]  54.4727   7.9922
[53,]  99.4141   4.8560
[54,]  89.1875   6.2305
[55,]  94.9883  23.7891
[56,]  95.7617   9.7070
[57,]  84.6250  14.7455
[58,]  97.4141   7.9023
```

```
[59,]  80.4375   9.4805
[60,] 104.3438   3.6016


# Os tamanhos das amostras devem ser informados:
n1<-30  # N° de observações referentes ao poço 1RJS0019RJ
n2<-30  # N° de observações referentes ao poço 3NA0002RJS

n<-n1+n2

# Informando que os dados devem ser vistos como matriz:
mdados<-matrix(mdados,nrow=n,ncol=2)

# Separando as duas amostras:
mdados1<-mdados[1:n1,]
n11<-n1+1
mdados2<-mdados[n11:n,]

mdados1<-matrix(mdados1,nrow=n1,ncol=2)
mdados2<-matrix(mdados2,nrow=n2,ncol=2)
mdados1
          [,1]    [,2]
 [1,]  84.1809  2.0278
 [2,]  86.7461  1.6348
 [3,]  89.6758  1.6054
 [4,]  89.6680  2.1108
 [5,]  85.0625  2.2480
 [6,]  92.8789  2.7148
 [7,]  84.7344  2.7734
 [8,]  94.3552  2.9285
 [9,]  89.8047  4.8696
[10,] 110.2070  9.6970
[11,]  93.5312  5.3892
[12,]  77.9180  2.7256
[13,]  86.4336  1.7183
[14,]  84.6033  1.7896
[15,]  91.7109  4.8125
[16,]  94.6211  4.9150
[17,]  76.0000  3.6597
```

```
[18,]  84.1450  2.0310
[19,]  86.8750  1.7886
[20,]  97.2070  6.4727
[21,]  83.5627  1.9921
[22,]  79.5117  2.1458
[23,]  79.3516  3.0627
[24,]  74.8906  4.2173
[25,]  79.6875 21.6758
[26,]  84.9375 32.1875
[27,]  98.4651 13.2656
[28,]  80.3125 33.9375
[29,]  78.5195  1.8826
[30,]  83.2959  2.0723
> mdados2
          [,1]     [,2]
 [1,]  91.2695   1.8425
 [2,]  90.9714   1.4395
 [3,]  95.6133   1.6209
 [4,]  85.5479   1.9060
 [5,]  90.6836   1.6438
 [6,]  92.8125   1.4741
 [7,]  95.5234   1.8086
 [8,] 102.0547   2.9922
 [9,]  96.8945   4.3594
[10,] 100.7148   2.5627
[11,]  93.0625   4.2539
[12,]  65.7500  18.1797
[13,]  86.2109   8.6641
[14,] 104.4895 274.0000
[15,]  86.3750  39.0611
[16,]  76.0117  44.6250
[17,]  96.9492 353.0000
[18,]  78.0625  10.8945
[19,]  87.8750   3.8750
[20,]  71.0391   7.2578
[21,]  99.6719   3.3254
[22,]  54.4727   7.9922
[23,]  99.4141   4.8560
[24,]  89.1875   6.2305
```

```
[25,]  94.9883  23.7891
[26,]  95.7617   9.7070
[27,]  84.6250  14.7455
[28,]  97.4141   7.9023
[29,]  80.4375   9.4805
[30,] 104.3438   3.6016

# Calculando os vetores de médias amostrais:
vma1<-apply(mdados1, 2, mean)
vma2<-apply(mdados2, 2, mean)
vma1
[1] 86.76311  6.14505
vma2
[1] 89.60759 29.23636

# Removendo os valores indisponíveis:
media.na<-function(w){mean(w,na.rm=T)}
covar<-function(A){var(A, na.rm = T)}

#-------------------------------------------------------------------------#
# 1° passo do procedimento "hold-out" de Lachenbruch: Fixar a matriz
# de dados referente à segunda amostra e variar a matriz de dados
# referente à primeira.
#-------------------------------------------------------------------------#

# Calculando a matriz de covariâncias amostrais para a segunda amostra
# (poço 3NA0002RJS), que está fixa:
S2<-covar(mdados2)
S2
         [,1]      [,2]
[1,] 133.2875  182.1375
[2,] 182.1375 6186.3401

# Calculando a nova matriz S1 (após a exclusão da i-ésima linha da matriz
# de dados referente à primeira amostra):

erro1<-0
correto1<-0
for (i in 1:n1)
```

```
{
A<-mdados1
x<-A[i,]
A[i,]<-NA
S1<-covar(A)

# Calculando a matriz de covariâncias amostrais combinadas (Spooled):
Sp<-((n1-2)*S1+(n2-1)*S2)/(n1+n2-3)

xbarra1<-apply(A, 2, media.na)
xbarra2<-vma2

# Computando o valor da função linear y e comparando-a com o ponto médio m:
y<- t(x-xbarra1)%*%solve(Sp)%*%(x-xbarra1)
m<- t(x-xbarra2)%*%solve(Sp)%*%(x-xbarra2)

ifelse(y<m,correto1<-correto1+1, erro1<-erro1+1)
}

#---------------------------------------------------------------------#
# 2° passo do procedimento "hold-out" de Lachenbruch: Fixar a matriz
# de dados referente à primeira amostra e variar a matriz de dados
# referente à segunda.
#---------------------------------------------------------------------#

# Calculando a matriz de covariâncias amostrais para a primeira amostra
# (poço 1RJS0019RJ), que está fixa:
S1<-covar(mdados1)

# Calculando a nova matriz S2 (após a exclusão da i-ésima linha da matriz
# de dados referente à segunda amostra):

erro2<-0
correto2<-0
for (i in 1:n2)
{
B<-mdados2
x<-B[i,]
B[i,]<-NA
```

```
S2<-covar(B)

# Calculando a matriz de covariâncias amostrais combinadas (Spooled):
Sp<-((n1-1)*S1+(n2-2)*S2)/(n1+n2-3)

xbarra1<-vma1
xbarra2<-apply(B, 2, media.na)

# Computando o valor da função linear y e comparando-a com o ponto médio m:
y<- t(x-xbarra1)%*%solve(Sp)%*%(x-xbarra1)
m<- t(x-xbarra2)%*%solve(Sp)%*%(x-xbarra2)

ifelse(y>=m,correto2<-correto2+1, erro2<-erro2+1)
}

#---------------------------------------------------------------------#

# Imprimindo os Resultados

# Matriz de Confundimento
cat(" ",fill=T)
cat("                  Matriz de Confundimento",fill=T)
cat(" ",fill=T)
cat("                                    Pertinência Prevista",fill=T)
cat("                                Poço 1RJS0019RJ  Poço 3NA0002RJS",fill=T)
cat("Pertinência   Poço 1RJS0019RJ    ",correto1,"     ",erro1,fill=T)
cat("   Real       Poço 3NA0002RJS    ",erro2,"      ",correto2,fill=T)
```

```
                  Matriz de Confundimento

                                      Pertinência Prevista
                                  Poço 1RJS0019RJ     Poço 3NA0002RJS
Pertinência     Poço 1RJS0019RJ         21                   9
   Real         Poço 3NA0002RJS         15                  15
```

```
#---------------------------------------------------------------------#
```

```
# Estimativa para a taxa de erro real:

E<-(erro1+erro2)/n
Ep<-100*E
cat(" ",fill=T)
cat(" Estimativa para a taxa de erro real .......",E,"=>",Ep,"%",fill=T)

Estimativa para a taxa de erro real ....... 0.4 => 40 %

# Nota-se que taxa de erro estimada é muito alta. Isso se deve ao fato
# da variância dos dados ser alta, principalmente em se tratando da
# variável ILD, em ambos os poços. Isso contribui para um aumento nas
# probabilidades de classificações erradas.
#-----------------------------------------------------------------------#
```

Os métodos estatísticos oferecem algumas vantagens nos estudos que envolvem avaliação de jazidas, como permitir a medição do erro que se comete ao fazer uma estimativa, possibilitando conhecer, de acordo com um grau de precisão desejado, a variação que se pode esperar em torno dos valores verdadeiros das variáveis analisadas. Outra vantagem adicional é a possibilidade de constatação da existência ou não de correlação entre as diferentes variáveis.

A aplicação da Estatística Clássica na estimativa de reservas está baseada na teoria das probabilidades, ou seja, no estudo das variáveis aleatórias independentes. Portanto, devemos enfatizar a importância de se utilizar dados cuja amostragem tenha sido feita de tal forma que concilie o caráter "independente" das variáveis geológicas em estudo com as características da natureza geológica, tendo em vista que estas variam de forma contínua ou descontínua de um ponto a outro, de acordo com a presença ou ausência de certos fenômenos naturais.

A aplicação destas técnicas na Indústria do Petróleo é bastante ampla, abrangendo avaliações de reservas através de análises de variáveis como teor de determinadas substâncias no subsolo, espessura das camadas, porosidade, permeabilidade e resistividade das rochas, dentre outras.

Capítulo 3

Métodos Geoestatísticos

3.1 Introdução

A indústria petrolífera tem sido estimulada a utilizar renovadas técnicas de caracterização de reservatório, pelos altos investimentos necessários no desenvolvimento de campos heterogêneos e pelo desejo de aumentar o fator de recuperação final das jazidas. Dentre essas técnicas, podemos citar as técnicas geoestatísticas, tópico especial da estatística aplicada cuja metodologia tem sido bem aceita, especialmente quando há a utilização de dados sísmicos tridimensionais.

A raízes da Geoestatística estão na indústria de minérios, na década de 50, quando o engenheiro de minas D. G. Krige e o estatístico H. S. Sichel desenvolveram novos métodos de estimação para reservas minerais espalhadas. Entre 1957 e 1962, o engenheiro francês G. Matheron, baseado nas observações de Krige, desenvolveu a Teoria das Variáveis Regionalizadas, a partir dos fundamentos da Geoestatística. Até 1968, a Geoestatística foi empregada para estimativas de reserva de hidrocarbonetos. Entre 1968 e 1970, foi desenvolvida a Teoria da Krigagem Universal (nome introduzido por Matheron em homenagem a Krige), para aplicação à cartografia submarina com tendência sistemática. Em 1972, Matheron criou a teoria Intrínseca de Ordem K, aplicada à meteorologia. Entre 1972 e 1973 surgiram os princípios da Análise Convexa, visando maximizar as reservas recuperáveis das jazidas subterrâneas. Em 1974 nasceu a teoria das funções de recuperação e, baseada nela, a Geoestatística não-linear aplicada

na seleção de reservas recuperáveis [10].

Uma, dentre as muitas vantagens da aplicação da Geoestatística, é o fato de ela necessitar e incentivar a interdisciplinaridade, assegurando uma maior troca de informações entre geólogos, engenheiros de petróleo, matemáticos e estatísticos e uma melhor interpretação da realidade geológica em estudo.

O objetivo da Geoestatística é melhorar as predições, através da construção de um modelo mais realista da heterogeneidade de um reservatório, usando métodos que não consideram médias, como propriedades de reservatório, importantes e assegurando que a realidade geológica não seja perdida durante a construção do modelo.

A Geoestatística produz numerosos resultados plausíveis, os quais requerem múltiplas simulações de escoamento em reservatório. Entretanto, os benefícios superam o tempo e custos adicionais.

3.2 Conceitos Básicos

Entenda-se por Variáveis Regionalizadas as variáveis cujos valores são relacionados de algum modo com a posição espacial onde os mesmos são obtidos, ou seja, é uma função que varia de um lugar para outro no espaço, com certa aparência de continuidade.

A continuidade atribuída às variáveis regionalizadas está relacionada com a variabilidade das propriedades da amostra com respeito à distância e direção, ou seja, com a tendência de tomarem valores mais próximos em dois pontos amostrados, quanto menos afastados geograficamente estejam os referidos pontos.

As variáveis regionalizadas são representadas, na prática, por uma certa quantidade de dados numéricos brutos disponíveis, a partir dos quais são obtidas informações sobre as características do fenômeno natural em estudo. Tais características são:

- **localização:** Uma variável regionalizada é numericamente definida por um valor, o qual está associado a uma amostra de tamanho, forma e orientação específicos. Essas características geométricas da amostra são denominadas suporte geométrico. O suporte geométrico não necessariamente compreende volumes, podendo se referir também a áreas e linhas. Vale salientar que somente no espaço geométrico onde a variável é susceptível de tomar valores definidos e no interior

do qual sua variação será estudada as variáveis regionalizadas tomam seus valores. Este espaço é denominado de campo geométrico, e pode, no nosso caso, ser uma parte ou todo o reservatório.

- **continuidade:** Dependendo do fenômeno observado, a variação espacial de uma variável regionalizada pode ser grande ou pequena. A existência de uma continuidade mais, ou menos, estável na variação de uma variável regionalizada pode ser expressa por meio de uma flutuação mais, ou menos, importante entre os valores de amostras vizinhas. Tal flutuação reflete o grau de dependência ou independência que existe entre um valor e outro. Quando essa continuidade é pouco definida e não pode ser confirmada, diz-se que há a presença do *efeito de pepita*. Quando os valores representativos das características do reservatório são totalmente independentes, trata-se de uma variável aleatória, considerada como um caso particular de variável regionalizada.

- **anisotropia ou zonalidade:** Fenômeno que indica se os valores da variável regionalizada não apresentam variações significativas ao longo de uma direção privilegiada, apresentando, por variações rápidas ou irregulares em outra(s) direção(ões).

3.3 Principais Objetivos da Geoestatística

Através das técnicas geoestatísticas são realizados estudos que levam em consideração a localização geográfica e a dependência espacial entre os dados, considerando, assim, as duas características essenciais das variáveis regionalizadas: o aspecto aleatório (já que os valores numéricos observados podem variar consideravelmente de um ponto a outro no espaço) e o aspecto espacial (visto que os valores numéricos observados não são inteiramente independentes), reproduzindo os fenômenos naturais, portanto, com maior fidelidade.

Desta forma, os dois principais objetivos de estudo da Geoestatística são:

- Tentar extrair, da aparente desordem dos dados disponíveis, uma imagem da variabilidade dos mesmos, e uma medida da correlação existente entre os valores

tomados em dois pontos do espaço. Este objetivo pode ser alcançado utilizando o variograma e está presente na análise estrutural.

- Medir a precisão de toda estimativa ou predição feita por meio de dados fragmentados, tornando necessária uma teoria de estimativa de reservas. Isto é feito usando a krigagem [10].

O variograma e a krigagem serão abordados mais detalhadamente nas sessões 3.5.2 e 3.5.3, respectivamente.

3.4 Funções Aleatórias

Para que sejam obtidos resultados satisfatórios, é muito importante o conhecimento, pelo menos parcial, da função densidade de probabilidade que governa a variável regionalizada, visto que o objetivo é estimar a variação da variável regionalizada em uma, duas ou três dimensões. Esse conhecimento pode ser baseado tanto num modelo teórico quanto numa análise empírica de uma amostra suficientemente grande. Devido à complexidade dessas variáveis regionalizadas, a alternativa da formulação de um modelo teórico é pouco utilizada, restando como solução a determinação empírica ou relativa das probabilidades presentes.

Seja Z uma variável aleatória. Um vetor aleatório a k componentes é definido pela sua função de distribuição

$$F(z_1, z_2, ..., z_k) = P(Z_1 \leq z_1, Z_2 \leq z_2, ..., Z_k \leq z_k), \tag{3.1}$$

onde $\boldsymbol{z} = (z_1, z_1, ..., z_k)$ é uma amostra da variável aleatória Z.

Seja $Z(\boldsymbol{x}_i)$ o valor de uma variável regionalizada Z obtido no ponto $\boldsymbol{x}_i$. Considerando que a função $F(z_1, z_2, ..., z_k)$ seja aleatória, supõe-se que a função de distribuição conjunta para quaisquer k componentes está bem definida [1].

$$F(Z_1, Z_2, ..., Z_k; \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k) = P(Z_1(\boldsymbol{x}_1) \leq z_1, Z_2(\boldsymbol{x}_2) \leq z_2, ..., Z_k(\boldsymbol{x}_k) \leq z_k).$$

A esperança de uma função aleatória é dada por

$$EZ(\boldsymbol{x}) = \int Z dF(Z, \boldsymbol{x}).$$

E a covariância por

$$Cov(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)) = E\{[Z(\boldsymbol{x}_i) - E(Z(\boldsymbol{x}_i))][Z(\boldsymbol{x}_j) - E(Z(\boldsymbol{x}_j))]\}.$$

Em se tratando de uma variável regionalizada, é possível realizar inferências estatísticas tomando-se por base apenas uma amostra, visto que a mesma é o resultado único de uma função casual. Este impasse é resolvido com a utilização da restrição estacionária, chamada de hipótese intrínseca, a qual permite o uso de resultados de uma variável regionalizada através da estimação pelo método dos momentos [4].

3.4.1 Funções Aleatórias Estacionárias

A função aleatória estacionária, caso particular de função aleatória, admite que todas as leis da função aleatória são invariantes para toda translação efetuada sobre os pontos $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$.

Sejam $Z(\boldsymbol{x}_i)$ e $Z(\boldsymbol{x}_i + \boldsymbol{h})$ dois valores de uma variável regionalizada Z obtidos nos pontos $\boldsymbol{x}_i$ e $\boldsymbol{x}_i + \boldsymbol{h}$, onde $\boldsymbol{h} = (h_1, h_2, ..., h_k)$ é um vetor com direção e orientação específica em um espaço de uma, duas ou três dimensões. Se a função aleatória é estacionária, temos que

$$F(Z_1, Z_2, ..., Z_k; \boldsymbol{x}_1 + \boldsymbol{h}, \boldsymbol{x}_2 + \boldsymbol{h}, ..., \boldsymbol{x}_k + \boldsymbol{h}) = F(Z_1, Z_2, ..., Z_k; \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k),$$

ou seja, para qualquer deslocamento $d = \mid \boldsymbol{h} \mid$, os dois primeiros momentos da diferença $[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{h})]$ não dependem da localização de Z, mas apenas de $\boldsymbol{h}$.

A diferença entre os valores $Z(\boldsymbol{x}_i)$ e $Z(\boldsymbol{x}_i + \boldsymbol{h})$ é outra variável casual.

A teoria das variáveis regionalizadas pressupõe que a variação de uma variável pode ser expressa pela soma de três componentes [2] e [6], a saber:

a) uma componente estrutural, associada a um valor médio constante ou a uma tendência constante;

b) uma componente aleatória, espacialmente correlacionada;

c) um ruído aleatório ou erro residual.

Se $\boldsymbol{x}$ representa uma posição em uma, duas ou três dimensões, então o valor da variável Z, em $\boldsymbol{x}$, é dada por [2]

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + \varepsilon^{'}(\boldsymbol{x}) + \varepsilon^{''},$$

onde

- $m(\boldsymbol{x})$ é uma componente determinística que descreve a parte estrutural de Z em $\boldsymbol{x}$;
- $\varepsilon^{'}(\boldsymbol{x})$ é um termo estocástico, que varia localmente e depende espacialmente de $m(\boldsymbol{x})$;
- $\varepsilon^{''}$ é um ruído aleatório não correlacionado, com distribuição normal de média zero e variância σ^2.

O primeiro passo na krigagem é definir uma função apropriada para a componente determinística $m(\boldsymbol{x})$. Para tanto, algumas hipóteses são necessárias [5] e [6].

3.4.2 Estacionariedade de Segunda Ordem

Uma função é denominada estacionária de segunda ordem quando a componente $m(\boldsymbol{x})$ é constante, ou seja, não há tendências na região. Desta forma, temos que

$$E\{Z(\boldsymbol{x})\} = E\{Z(\boldsymbol{x}+\boldsymbol{h})\} = m(\boldsymbol{x}) = m, \tag{3.2}$$

isto é, a diferença média entre os valores observados em $\boldsymbol{x}$ e $\boldsymbol{x}+\boldsymbol{h}$, separados por um vetor de distância $\boldsymbol{h}$ (módulo e direção) é nula:

$$E\{Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\} = 0.$$

Além disso, também é admitido que a covariância entre os pares $Z(\boldsymbol{x})$ e $Z(\boldsymbol{x}+\boldsymbol{h})$, separados por um vetor distância $\boldsymbol{h}$, existe e depende somente de $\boldsymbol{h}$.

Com isso, temos:

$$\begin{aligned} C(\boldsymbol{h}) &= Cov[Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})] \\ &= E\{[Z(\boldsymbol{x}) - m][Z(\boldsymbol{x}+\boldsymbol{h}) - m]\} \\ &= E\{Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h}) - mZ(\boldsymbol{x}+\boldsymbol{h}) - mZ(\boldsymbol{x}) + m^2\} \\ &= E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] - mE[Z(\boldsymbol{x}+\boldsymbol{h})] - mE[Z(\boldsymbol{x})] + m^2. \end{aligned}$$

Da equação (3.2) temos que:

$$\begin{aligned} C(\boldsymbol{h}) &= E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] - m^2 - m^2 + m^2 \\ &= E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] - m^2, \quad \forall \boldsymbol{x}. \end{aligned} \tag{3.3}$$

Ou seja,

$$E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] = C(\boldsymbol{h}) + m^2, \quad \forall \boldsymbol{x}. \tag{3.4}$$

Na equação (3.3), a estacionariedade da covariância implica na estacionariedade da variância:

$$\begin{aligned} Var[Z(\boldsymbol{x})] &= E\{[Z(\boldsymbol{x}) - m]^2\} \\ &= E\{Z^2(\boldsymbol{x}) - 2mZ(\boldsymbol{x}) + m^2\} \\ &= E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{0})] - 2m^2 + m^2 \\ &= E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{0})] - m^2 \\ &= C(\boldsymbol{0}), \quad \forall \boldsymbol{x}. \end{aligned} \tag{3.5}$$

Além disso, a estacionariedade da covariância também implica na estacionariedade do variograma, definido por:

$$\begin{aligned} 2\gamma(\boldsymbol{h}) &= E\{[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})]^2\} \\ &= E\{Z^2(\boldsymbol{x}) - 2Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h}) + Z^2(\boldsymbol{x}+\boldsymbol{h})\} \\ &= E[Z^2(\boldsymbol{x})] - 2E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] + E[Z^2(\boldsymbol{x}+\boldsymbol{h})]. \end{aligned} \tag{3.6}$$

De (3.3) temos que:

$$E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] = C(\boldsymbol{h}) + m^2, \tag{3.7}$$

e de (3.5) temos que:

$$E[Z^2(\boldsymbol{x})] = E[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{0})] = C(\boldsymbol{0}) + m^2. \tag{3.8}$$

Substituindo as equações (3.7) e (3.8) na equação (3.6), obtém-se:

$$2\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) + m^2 - 2[C(\boldsymbol{h}) + m^2] + E[Z^2(\boldsymbol{x}+\boldsymbol{h})], \tag{3.9}$$

e como $E[Z^2(\boldsymbol{x})] = E[Z^2(\boldsymbol{x}+\boldsymbol{h})]$, temos que

$$\begin{aligned} 2\gamma(\boldsymbol{h}) &= C(\boldsymbol{0}) + m^2 - 2[C(\boldsymbol{h}) + m^2] + C(\boldsymbol{0}) + m^2 \\ &= 2C(\boldsymbol{0}) - 2C(\boldsymbol{h}), \end{aligned} \tag{3.10}$$

de onde segue que

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}). \tag{3.11}$$

A função $\gamma(\boldsymbol{h})$ é conhecida na teoria das variáveis regionalizadas como semivariograma.

Da relação (3.11) temos que a covariância e o semivariograma são formas alternativas de caracterizar a autocorrelação dos pares $Z(\boldsymbol{x})$ e $Z(\boldsymbol{x}+\boldsymbol{h})$ e separados pelo vetor $\boldsymbol{h}$, sob a hipótese de estacionariedade de segunda ordem.

Como a hipótese de estacionariedade de segunda ordem supõe a existência de uma covariância e, conseqüentemente, de uma variância finita, equação (3.5), o correlograma $Y(\boldsymbol{h})$ pode ser definido como:

$$Y(\boldsymbol{h}) = \frac{C(\boldsymbol{h})}{C(\boldsymbol{0})}. \tag{3.12}$$

Da relação (3.11), temos que

$$C(\boldsymbol{h}) = C(\boldsymbol{0}) - \gamma(\boldsymbol{h}). \tag{3.13}$$

Substituindo (3.13) em (3.12), temos:

$$\begin{aligned} Y(\boldsymbol{h}) &= \frac{C(\boldsymbol{0}) - \gamma(\boldsymbol{h})}{C(\boldsymbol{0})} \\ &= 1 - \frac{\gamma(\boldsymbol{h})}{C(\boldsymbol{0})} \end{aligned} \tag{3.14}$$

As hipóteses de estacionariedade de segunda ordem, ou seja, $\exists C(\boldsymbol{h}) \Rightarrow \exists Var[Z(\boldsymbol{x})] = C(\boldsymbol{0})$ e $\exists C(\boldsymbol{h}) \Rightarrow \exists \gamma(\boldsymbol{h})$, podem não ser satisfeitas para alguns fenômenos físicos que apresentam uma capacidade infinita de dispersão [6] pois, este caso implica a não existência de $C(\boldsymbol{h})$ e de $Var[Z(\boldsymbol{x})]$, podendo existir, entretanto, $\gamma(\boldsymbol{h})$. Para tais situações, uma hipótese menos restritiva, a hipótese intrínseca, pode ser aplicável.

3.4.3 Estacionariedade Intrínseca

Como na hipótese anterior, aqui se admite também que $E[Z(\boldsymbol{x})] = m(\boldsymbol{x}) = m, \quad \forall \boldsymbol{x}$.

Além disso, admite-se que a variância das diferenças depende somente do vetor distância $d = \mid \boldsymbol{h} \mid$, isto é:

$$Var[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})] = E\{[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})]^2\} = 2\gamma(\boldsymbol{h}). \tag{3.15}$$

Por ser a menos restritiva e requerer apenas a existência e estacionariedade do variograma, sem nenhuma restrição quanto à existência de variância finita, esta hipótese é a mais freqüentemente admitida em Geoestatística [6].

Em se tratando das hipóteses da Krigagem Universal, admite-se que $m(\boldsymbol{x})$ é a tendência principal (*drift*) e que $C(\boldsymbol{h})$ e $\gamma(\boldsymbol{h})$ possuem estacionariedade dentro de uma vizinhança de tamanho restrito. Além disso, supõe-se que $E[Z(\boldsymbol{x})] = m(\boldsymbol{x})$, onde $m(\boldsymbol{x})$ deixa de ser estacionária, variando de modo regular dentro de tal vizinhança. Não somente a covariância e o variograma são definidos a partir de valores experimentais, mas também o tamanho da vizinhança onde as hipóteses mantêm-se válidas [6].

3.5 Elementos Básicos para um Estudo Geoestatístico

3.5.1 Extração de Dados

Análises de dados em estatística clássica incluem computação de médias, variâncias e outras medidas descritivas, diagramas de dispersão para o estudo da relação entre duas ou mais variáveis e identificação de sub-populações e pontos de alavanca. Apenas após a reunião e descrição dos dados, as análises podem ser feitas com mais segurança, evitando resultados incoerentes. Outro fator que torna a organização dos dados imprescindível é a susceptibilidade de erros cometidos nos estudos envolvendo grandes volumes de dados e utilização de computadores.

3.5.2 Modelagem e Análise de Continuidade Espacial

Na indústria de Petróleo, para uma eficiente produção de hidrocarbonetos, é necessário entender as escalas e os aspectos direcionais das características físicas das propriedades das rochas-reservatório, bem como as características do modelo espacial associado às variáveis, tais como porosidade, saturação, etc, frutos de um vasto número de processos químicos e físicos bastante complexos. A componente espacial torna estas variáveis complicadas, sendo necessário o reconhecimento das incertezas na identificação da distribuição das mesmas entre os poços.

O modelo precisa, então, descrever a continuidade, a anisotropia e as propriedades azimutais dos dados. As análises de continuidade espacial quantificam a variabilidade das propriedades da amostra relacionadas com a distância e direção, comparando valores de dados em uma locação com valores do mesmo atributo em outras locações. Tais

análises, geralmente envolvem grandes volumes de dados e utilização de computadores, através de *softwares* especiais.

Uma medida muito comum de continuidade espacial é o variograma.

Variograma

Como mencionado na seção (3.3), uma abordagem mais detalhada do variograma será feita a seguir.

Sejam X e Y duas variáveis regionalizadas, onde $X = Z(\boldsymbol{x})$ e $Y = Z(\boldsymbol{x} + \boldsymbol{h})$, referentes ao mesmo atributo (por exemplo, o teor de um contaminante no solo), medido em duas posições diferentes, conforme ilustra a figura a seguir, onde $\boldsymbol{x}$ denota uma posição em duas dimensões, com componentes (x_1, y_1), e $\boldsymbol{h}$ um vetor distância (módulo e direção) que separa os pontos:

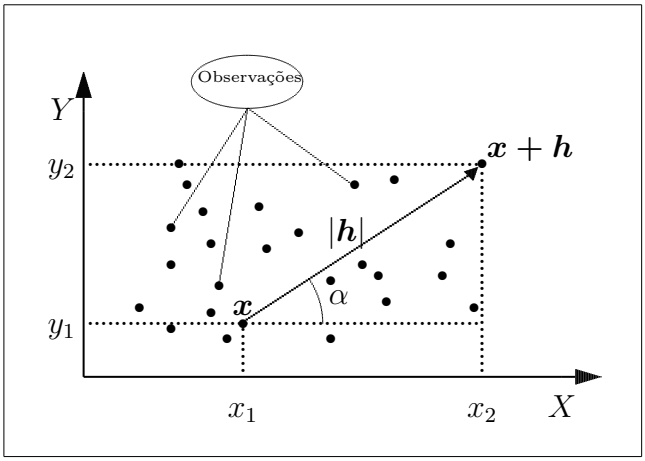


Figura 3.1: Localizações dos pontos.

O variograma $2\gamma(\boldsymbol{h})$, definido como sendo a esperança matemática do quadrado da diferença entre os valores de pontos no espaço, separados por $d = \mid \boldsymbol{h} \mid$, representa o nível de dependência entre as duas variáveis regionalizadas, X e Y.

$$\begin{aligned} 2\gamma(\boldsymbol{h}) &= E\{[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})]^2\} \\ &= Var[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})]. \end{aligned} \tag{3.16}$$

Através de uma amostra $Z(\boldsymbol{x}_i), i = 1, 2, ..., n$, o variograma pode ser estimado por

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{h})]^2, \tag{3.17}$$

onde:

- $2\widehat{\gamma}(\boldsymbol{h})$: variograma estimado;

- $N(\boldsymbol{h})$: número de pares de valores medidos, $Z(\boldsymbol{x}_i)$ e $Z(\boldsymbol{x}_i+\boldsymbol{h})$, separados por uma distância $d =\mid \boldsymbol{h} \mid$;

- $Z(\boldsymbol{x}_i)$ e $Z(\boldsymbol{x}_i+\boldsymbol{h})$: valores da i-ésima observação da variável regionalizada, coletados nos pontos $\boldsymbol{x}_i$ e $\boldsymbol{x}_i+\boldsymbol{h}$, $i = 1, 2, ..., n$, separados por $d =\mid \boldsymbol{h} \mid$.

Alguns autores definem variograma considerando o que comumente se refere como semivariograma, termo este advindo da divisão por dois para compatibilização da fórmula

$$\gamma(\boldsymbol{h}) = \frac{1}{2}2\gamma(\boldsymbol{h}). \tag{3.18}$$

De (3.17), temos que a função do semivariograma pode ser estimada por:

$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{h})]^2. \tag{3.19}$$

Parâmetros do Semivariograma

Quando os valores de dados em uma localização são comparados com valores do mesmo atributo em outras localizações, espera-se que observações geograficamente mais próximas tenham um comportamento mais semelhante entre si do que aquelas separadas por distâncias maiores. Em outras palavras, espera-se que as diferenças $[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{h})]$ sejam reduzidas, à medida que a distância que os separa, $d =\mid \boldsymbol{h} \mid$, decresça, ou seja, que $\gamma(\boldsymbol{h})$ aumente com a distância $d =\mid \boldsymbol{h} \mid$.

A figura a seguir ilustra o exemplo de um semivariograma experimental com características muito próximas do ideal [22].

Os parâmetros do semivariograma podem ser observados diretamente da figura (3.2).

- *Alcance* (a): distância abaixo da qual as amostras apresentam-se correlacionadas espacialmente. Na figura (3.2), o alcance ocorre próximo de 25.

- *Patamar* (C): é o valor do semivariograma correspondente a seu alcance (a), ou seja, é o valor constante atingido por $\hat{\gamma}(\boldsymbol{h})$ quando a distância entre os dados, d, cresce. Deste ponto em diante, considera-se que não existe mais dependência espacial entre as amostras. Na figura (3.2), o patamar é aproximadamente 1,75.

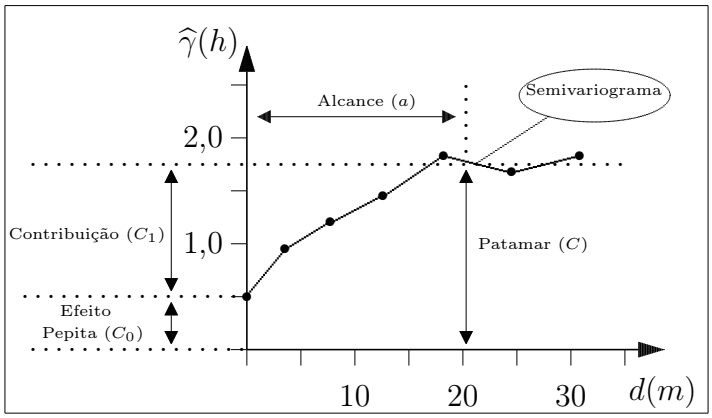


Figura 3.2: Exemplo de Semivariograma.

- *Efeito Pepita* (C_0): por definição, das equações (3.16) e (3.18), temos que $\gamma(\boldsymbol{h}) = 0$. Na prática, entretanto, à medida que $\boldsymbol{h} \to 0$, $\gamma(\boldsymbol{h})$ se aproxima de um valor positivo denominado *Efeito Pepita* (C_0). O valor de (C_0) revela a descontinuidade do semivariograma para distâncias menores do que a menor distância entre as amostras. Parte desta descontinuidade pode ser também devida a erros de medição [11], mas é impossível quantificar se a maior contribuição provém dos erros de medição ou da variabilidade de pequena escala, não captada pela amostragem. Na figura 3.2, o efeito pepita é próximo de 0,5.

- *Contribuição* (C_1): é a diferença entre o patamar (C) e o Efeito Pepita (C_0). Na figura 3.2, a contribuição é de, aproximadamente, 1,25.

Existem três tipos de variograma, a saber:

- variograma teórico: é o variograma de referência;

- variograma experimental: é o variograma obtido a partir do conjunto de amostras derivadas da amostragem realizada;

- variograma verdadeiro: é o variograma real do reservatório.

O principal objetivo de um estudo estrutural consiste em verificar qual é o variograma teórico que melhor se ajusta ao variograma experimental, de tal forma que o variograma verdadeiro possa ser inferido a partir do modelo teórico citado. Após a identificação do variograma teórico adequado, suas propriedades são tomadas como base para a análise variográfica e para a avaliação das reservas.

3.5.3 Validação do Modelo

Com o objetivo de testar a eficiência do modelo encontrado, os valores estimados são comparados com os valores observados, através, por exemplo, do histograma dos erros de estimação padronizados, o qual corresponde aos valores estimados menos os valores observados, divididos pela variância da krigagem. Se o histograma for simétrico em torno da média 0 (zero), as estimativas são não tendenciosas.

Krigagem

A krigagem é um processo de estimação de valores de variáveis regionalizadas, a partir de valores adjacentes enquanto considerados independentes na análise variográfica [14]. Por meio dela, pode-se obter:

- A previsão do valor pontual de uma variável regionalizada em um local específico dentro do espaço geométrico (trata-se de um procedimento exato de interpolação que leva em consideração todos os valores observados);
- O cálculo médio de uma variável regionalizada para um volume maior do que o suporte geométrico;
- A estimação da tendência principal (*drift*), de modo similar à superfície de tendência.

A diferença entre a krigagem e outros algoritmos à disposição é que ela fornece, além dos valores estimados, o erro associado a tal estimação. Além disso, a maneira como os pesos são atribuídos às distintas amostras também são diferentes. No caso de interpolação linear simples, por exemplo, os pesos são todos iguais a $1/N$, onde N = número de amostras. Na interpolação baseada no inverso do quadrado das distâncias, os pesos são definidos como o inverso do quadrado da distância que separa o valor interpolado dos valores observados. Na krigagem, o procedimento é semelhante ao de interpolação por média móvel ponderada, exceto que aqui os pesos ótimos a serem associados às amostras que irão fornecer estimativas em um ponto, uma área ou um volume são determinados a partir de uma análise espacial, baseada no semivariograma experimental. Como o semivariograma é uma função da distância entre locais de amostragens, mesmo mantendo-se o mesmo número de amostras, os pesos serão

diferentes, de acordo com seu arranjo geográfico. Além disso, a krigagem fornece, em geral, estimativas não tendenciosas e com variância mínima.

Krigagem Simples

Seja Z uma propriedade do solo, observada em n pontos distintos, com coordenadas representadas pelo vetor $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$, onde $\boldsymbol{x}_i$ identifica uma posição em duas dimensões, representada pelos pares de coordenadas (x_{i1}, x_{i2}), para $i = 1, 2, ..., n$. Assim, tem-se um conjunto de valores $Z(\boldsymbol{x}_i)$, $i = 1, 2, ..., n$. Suponha que o objetivo seja estimar o valor desconhecido de Z no ponto $\boldsymbol{x}_0$, $Z(\boldsymbol{x}_0)$. Este pode ser estimado a partir de uma combinação linear dos valores observados, adicionado a um parâmetro λ_0 [13]

$$Z_{\boldsymbol{x}_0}^{*} = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i). \tag{3.20}$$

Então, temos:

$$\begin{aligned} E\big[Z_{\boldsymbol{x}_0}^{*}\big] &= E\big[\lambda_0 + \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i)\big] \\ &= \lambda_0 + \sum_{i=1}^{n} \lambda_i E\big[Z(\boldsymbol{x}_i)\big]. \end{aligned} \tag{3.21}$$

Deseja-se um estimador não tendencioso, isto é,

$$E\big[Z_{\boldsymbol{x}_0} - Z_{\boldsymbol{x}_0}^{*}\big] = 0, \tag{3.22}$$

ou seja,

$$E\big[Z_{\boldsymbol{x}_0}\big] = E\big[Z_{\boldsymbol{x}_0}^{*}\big]. \tag{3.23}$$

Substituindo a equação (3.21) em (3.23), obtemos o parâmetro λ_0

$$\lambda_0 = E\big[Z_{\boldsymbol{x}_0}\big] - \sum_{i=1}^{n} \lambda_i E\big[Z(\boldsymbol{x}_i)\big]. \tag{3.24}$$

Substituindo o valor de λ_0 na equação (3.20), obtém-se o estimador

$$Z_{\boldsymbol{x}_0}^{*} = E\big[Z_{\boldsymbol{x}_0}\big] - \sum_{i=1}^{n} \lambda_i E\big[Z(\boldsymbol{x}_i)\big] + \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i). \tag{3.25}$$

O método de krigagem simples supõe que a média (m) é conhecida e constante *a priori*, então

$$E\big[Z_{\boldsymbol{x}_0}\big] = E\big[Z(\boldsymbol{x}_i)\big] = m. \tag{3.26}$$

Substituindo a equação (3.26) em (3.25), o estimador de krigagem simples fica

$$\begin{aligned} Z^{*}_{\boldsymbol{x}_0} &= m - \sum_{i=1}^{n}\lambda_i m + \sum_{i=1}^{n}\lambda_i Z(\boldsymbol{x}_i) \\ &= m + \sum_{i=1}^{n}\lambda_i \big[Z(\boldsymbol{x}_i) - m\big]. \end{aligned} \tag{3.27}$$

Minimizando a variância do erro $Var\big[Z_{\boldsymbol{x}_0} - Z^{*}_{\boldsymbol{x}_0}\big]$, os pesos λ_i são obtidos a partir do seguinte sistema de equações, denominado sistema de krigagem simples [13]:

$$\sum_{j=1}^{n}\lambda_i C(\boldsymbol{x}_i, \boldsymbol{x}_j) = C(\boldsymbol{x}_i, \boldsymbol{x}_0), \text{ para } i = 1, 2, ..., n, \tag{3.28}$$

onde

- $C(\boldsymbol{x}_i, \boldsymbol{x}_j)$ refere-se à função de covariância correspondente a um vetor $\boldsymbol{h}$, com origem em $\boldsymbol{x}_i$ e extremidade em $\boldsymbol{x}_j$.
- $C(\boldsymbol{x}_i, \boldsymbol{x}_0)$ refere-se à função de covariância correspondente a um vetor $\boldsymbol{h}$, com origem em $\boldsymbol{x}_i$ e extremidade no ponto $\boldsymbol{x}_0$ a ser estimado.

Por exemplo, para $n = 2$, o sistema de krigagem simples constitui-se de duas equações a duas incógnitas (λ_1, λ_2), a saber:

$$\begin{cases} \lambda_1 C_{11} + \lambda_2 C_{12} = C_{10} \\ \lambda_1 C_{21} + \lambda_2 C_{22} = C_{20}. \end{cases} \tag{3.29}$$

A correspondente variância mínima do erro, denominada variância de krigagem simples (σ^2_{KS}), é dada por [13]

$$\begin{aligned} \sigma^2_{KS} &= Var\big[Z_{\boldsymbol{x}_0} - Z^{*}_{\boldsymbol{x}_0}\big] \\ &= C(\mathbf{0}) - \sum_{i=1}^{n}\lambda_i C(\boldsymbol{x}_i, \boldsymbol{x}_0). \end{aligned} \tag{3.30}$$

Em notação matricial, o sistema de krigagem simples é escrito como:

$$\boldsymbol{K}\boldsymbol{\lambda} = \boldsymbol{k} \Rightarrow \boldsymbol{\lambda} = \boldsymbol{K}^{-1}\boldsymbol{k}, \tag{3.31}$$

onde $\boldsymbol{K}$ e $\boldsymbol{k}$ são as matrizes das covariâncias e $\boldsymbol{\lambda}$ o vetor dos pesos, com

$$\boldsymbol{K} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix}, \ \boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} \text{ e } \boldsymbol{k} = \begin{bmatrix} C_{10} \\ C_{20} \\ \vdots \\ C_{n0} \end{bmatrix}. \tag{3.32}$$

A variância de krigagem simples é dada por [13]

$$\sigma^2_{KS} = C(\mathbf{0}) - \boldsymbol{\lambda}^T \boldsymbol{k}.$$

A krigagem é um método determinístico que tem solução única, oferecendo a melhor estimativa, no sentido de levar em consideração a localização espacial dos pontos. No entanto, ela não representa a variabilidade atual do atributo estudado. Por este motivo, abordaremos a seguir o conceito de modelagem estocástica e como esta é usada para modelar a heterogeneidade do reservatório de maneira mais realista.

3.5.4 Simulação Estocástica

Diferentemente da krigagem, os métodos estocásticos são capazes de contemplar os detalhes capturados utilizando a variância da estimação, objeto de grande interesse dos geocientistas e engenheiros de reservatório.

Sabemos que os processos naturais responsáveis pela formação de reservatórios não são puramente aleatórios [4]. Por este motivo, os métodos estocásticos são geralmente pouco utilizados. No entanto, devemos levar em consideração que apesar dos reservatórios não serem frutos de processos aleatórios, eles possuem atributos que os fazem se comportar como se o fossem. Por exemplo, podemos citar os processos químicos e físicos, os quais mudam as características do reservatório desde o seu estado inicial, confundindo os resultados predições, mesmo quando os processos são compreendidos. Estas mudanças resultam em um comportamento que pode ser capturado usando princípios estocásticos.

Além disso, a modelagem estocástica fornece aos geocientistas e engenheiros de reservatório muitos modelos equiprováveis, denominados de realizações. A solução obtida com a krigagem é a média de numerosas realizações, e a variabilidade entre os diferentes resultados é a medida de incerteza em qualquer locação. Assim, o desvio padrão de todos os valores simulados em uma locação fixa é uma quantificação da sua incerteza.

Muitas são as razões da utilização da simulação estocástica. Algumas delas serão abordadas a seguir.

Captura da Heterogeneidade

Cada vez mais se torna aparente a maior precisão das predições de performance de um reservatório, quando estas são baseadas em modelos que refletem a possível heterogeneidade do mesmo, pois a heterogeneidade tem um grande impacto nas características do escoamento. Como a simulação estocástica preserva a variância dos dados observados, as realizações são mais irregulares em aparência, mostrando mais variabilidade nos espaços entre os poços, tornado-se mais próximas das condições reais. Apesar de cada realização estocástica resultar em uma figura diferente do escoamento do fluido, cada realização geralmente fornece uma informação mais realista sobre o comportamento atual do fluxo do que a fornecida pelos modelos determinísticos convencionais. Apesar de serem criadas muitas realizações estocásticas para entender e quantificar as incertezas, apenas uma realização é usada para simulação e predição de performance.

Simulação das Fácies e/ou das Propriedades da Rocha

Durante a construção de um modelo estocástico para um reservatório, alguns passos devem ser seguidos.

A observação da arquitetura do reservatório, geralmente primeira prioridade, consiste da análise da estrutura completa dos elementos: falhas, topo e base do reservatório, etc.

O segundo passo é identificar diferentes unidades geológicas, baseadas em uma seqüência de princípios estratigráficos.

O passo seguinte envolve a modelagem da distribuição espacial de fácies deposicionais ou de litofácies. As informações das fácies deposicionais geralmente fornecem maior clareza em termos de geometrias espaciais, porém não são obtidas com facilidade. As litofácies são mais facilmente obtidas, mas não garantem respeitar os limites das fácies deposicionais atuais. No entanto, as duas são fortemente relacionadas, e é comum neste ponto o agrupamento de fácies que exibam similaridades nas propriedades petrofísicas e de saturação dentro de unidades mais gerais, chamadas de litotipos. Os litotipos são, então, modelados e os resultados devem refletir a acomodação das unidades de escoamento.

O passo final na construção do modelo geológico de alta resolução é propagar os

litotipos com a rocha e as propriedades do fluido.

Respeito e Interação das Informações Complexas

Os métodos estocásticos, assim como a krigagem, também permitem a incorporação de uma ampla escala de informação a qual não pode ser acomodada pela maioria dos métodos convencionais. Muitos estão interessados em simulação estocástica, não pela abrangência de seus resultados plausíveis, mas por sua habilidade de integrar simultaneamente os dados *soft*, como, por exemplo, testes sísmicos ou de poços, visto que esta integração aumenta a segurança das previsões em locações fora dos pontos de controle, onde estão disponíveis apenas as formas secundárias dos dados.

Acesso às Incertezas

Quem faz predições a respeito da performance de um reservatório, entende que sempre existe incerteza no seu modelo. Predições de performance ou predições volumétricas são, muitas vezes, baseadas no modelo do "melhor" caso. Entretanto, os pesquisadores também podem estar interessados em outros modelos, como os casos pessimistas e os casos otimistas. Um mínimo de três modelos permite a eles verificar se o plano de desenvolvimento, baseado no cenário do "melhor" caso, é flexível o bastante para tratar de uma escala de incerteza.

A simulação estocástica oferece muitos modelos consistentes com os dados de entrada. Um aspecto crítico é a crença em algum "espaço de incerteza" amostrado inviesadamente e adequadamente por um conjunto de realizações. Os resultados podem, então, ser resumidos mais como uma distribuição de probabilidade do que como uma simples seleção de algum número limitado de realizações plausíveis de um grande conjunto de dados.

Depois de simuladas as imagens, deve-se examinar qual(ais) delas é(são) geologicamente aceitável(eis). Cada uma das imagens simuladas deve ser examinada para determinar se é uma representação razoável do que é conhecido a respeito do reservatório. A(s) imagem(ns) simulada(s) que não seguir(em) esse comportamento deve(m) ser descartada(s).

Como a modelagem estocástica gera realizações independentes, os numerosos resultados são freqüentemente reprocessados para quantificar as incertezas. Mapas

plausíveis gerados de um acompanhamento de imagens simuladas incluem o desvio padrão, que é usado para analisar incerteza, a incerteza ou risco, a isoprobabilidade e a média das n simulações condicionais, onde, em cada célula, o programa computa o valor médio baseado nos valores de todas as simulações naquela locação e, além disso, quando n é grande, o mapa converge para a solução obtida por krigagem.

A Simulação Estocástica

A simulação estocástica é uma técnica designada para respeitar os dados medidos, reproduzir o histograma dos dados, respeitar o modelo espacial, ser consistente com os dados secundários, e acessar as incertezas no modelo do reservatório.

Vários métodos de simulação estão disponíveis. A escolha de um deles depende da meta a ser alcançada e dos tipos e disposição dos dados. Comumente, os métodos utilizados são *turning bands*, simulação seqüencial, *simulated annealing*, campo de probabilidade, decomposição de matrizes e Boolean (ou modelos de *object-based* tais como o processo de *marked-point*). Neste trabalho, receberão maior ênfase os métodos de simulação seqüencial, campo de probabilidade e de decomposição matricial.

(i) **Simulação Seqüencial**

Segue o seguinte processo:

- Seleção aleatória de um nó da grade ainda não simulado;
- Utilização da krigagem para computar a função distribuição de probabilidade cumulativa local ($lcpd$), com média zero e variância unitária. A computação da ($lcpd$) depende do método de simulação usado;
- Exclusão aleatoriamente de um único valor, z_i, da ($lcpd$), cuja extensão máxima seja duas vezes o desvio padrão em torno da média, m_i;
- Criação de um novo valor simulado $ZS_i^* = m_i + z_i$;
- Inclusão de ZS_i^* no conjunto dos dados condicionais, assegurando que valores em espaços próximos tenham uma correta escala de correlação;
- Repetição dessas etapas até que todos os nós da grade tenham um valor simulado.

O processo de seleção é aleatório, mas pode ser repetido. Para cada simulação, os nós da grade são misturados em uma ordem definida por um valor semente. Cada semente aleatória corresponde a uma única ordem dos nós da grade, e diferentes valores semente produzem diferentes caminhos dentro da grade. Apesar do número total de disposições ser possivelmente muito grande, cada caminho aleatório é unicamente identificado e repetido.

(ii) **Campo de Probabilidade**

Este método é uma alternativa para os métodos de simulação seqüenciais. Na simulação seqüencial, o valor retirado da distribuição de probabilidade cumulativa local (*lcpd*) em um particular nó da grade é tratado como dado *hard* e incluído como dado condicional local. Isto assegura que valores espacialmente próximos tenham a correta escala de correlação. De outra forma, a imagem simulada poderia conter altas escalas de variabilidade. A idéia utilizada pela simulação por campo de probabilidade, ou *P-field*, é aumentar a eficiência de computação da (*lcpd*) dos dados originais de poços. A simulação *P-field* tenta resolver o problema das altas escalas de variabilidade através do controle da amostragem das distribuições, em detrimento do controle das distribuições como na simulação seqüencial. A vantagem é a não necessidade de recomputação da (*lcpd*) nas realizações, aumentando de forma gigantesca a velocidade de computação.

(iii) **Decomposição Matricial**

Alguns métodos de simulação envolvem decomposição matricial. A decomposição LU, por exemplo, utiliza diferentes resultados criados por multiplicação de vetores aleatórios por uma matriz pré-calculada, derivada da informação de continuidade espacial (tipicamente de um variograma ou um correlograma). Métodos matriciais podem ser vistos como simulação seqüencial porque a multiplicação das linhas da matriz pré-calculada pelo vetor aleatório pode ser construído como um processo seqüencial em que o valor de cada nó sucessivo depende dos valores dos nós previamente simulados. Apenas centenas de nós podem ser simulados ao mesmo tempo usando métodos de decomposição de matrizes. Os modelos amplos são simulados através da união de modelos menores.

Capítulo 4

Aplicação a Dados da Indústria de Petróleo e Gás

4.1 Introdução

A necessidade de identificar e quantificar os parâmetros geológicos que caracterizam um reservatório petrolífero é antiga e ficou bem evidente na década de 60, quando foram efetuados estudos, buscando fazer associação de dados de afloramento com dados de testemunho e perfis geofísicos de poços [18].

A preocupação em identificar os parâmetros do reservatório que mais interferem no escoamento de fluidos e modelá-los numa escala compatível com o estudo de simulação de fluxo também é antiga. Desde a década de 70 sentiu-se a necessidade de descrever adequadamente esses parâmetros e vários modelos e escalas de heterogeneidade foram propostos com o objetivo de definir as heterogeneidades, como também as incertezas inerentes ao conhecimento do reservatório. Em particular, os modelos da geoestatística baseiam-se, em parte, na teoria das probabilidades, reconhecendo e incorporando as incertezas advindas do processo de obtenção dos dados.

A bacia de Campos, localizada no estado do Rio de Janeiro, é atualmente a bacia *offshore* mais produtiva do Brasil. Nela está situado o Campo Escola de Namorado, onde foram perfurados 56 poços, dentre os quais 36 são produtores, e 19 foram testemunhados.

Diversos dados obtidos a partir destes poços podem ser encontrados em um CD,

cedido pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Dentre estes dados, encontram-se alguns dos perfis, mencionados no Capítulo 1, como pode ser observado na Figura 4.1. Neste trabalho, optamos por utilizar apenas dados de perfis de oito poços do Campo de Namorado, sendo um poço pioneiro e sete poços de extensão, e em algumas profundidades arbitrariamente escolhidas, tendo em vista o grande volume de dados coletados, bem como a necessidade de se resguardar parte dos dados para que estes possam ser utilizados na validação dos resultados. O *software* utilizado no desenvolvimento desta pesquisa foi o R, por ser um *software* gratuito muito rico em ferramentas estatísticas e que pode ser encontrado no site http://www.r-project.org.

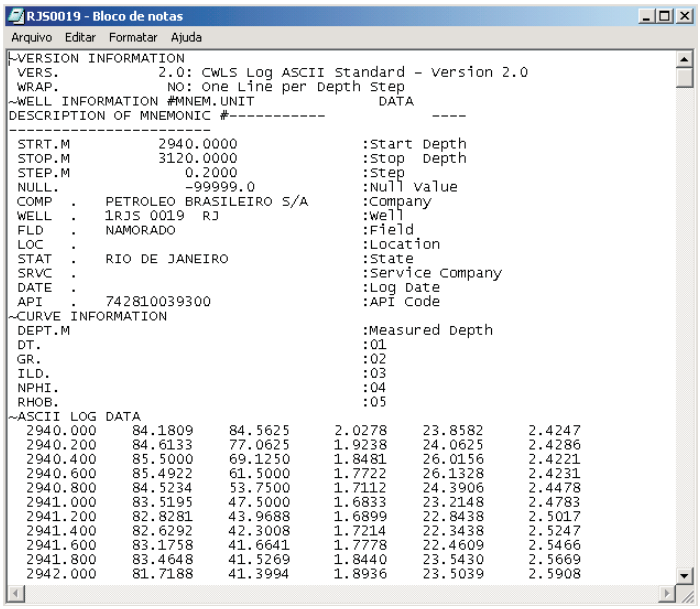


Figura 4.1: Dados no formato original.

4.2 Transformação das Coordenadas dos Poços

A localização dos poços, cujos dados estão contidos no CD anteriormente mencionado, precisou passar por uma prévia modificação, tendo em vista que estas foram apenas disponibilizadas em latitude e longitude. Assim, desenvolvemos uma pequena rotina, através da qual as unidades dos dados foram transformadas em coordenadas x e y, com relação à Linha do Equador e ao Meridiano de Greenwhich. Inicialmente, os dados referentes aos oito poços foram dispostos no formato utilizado pelo R, conforme

Figura 4.2.

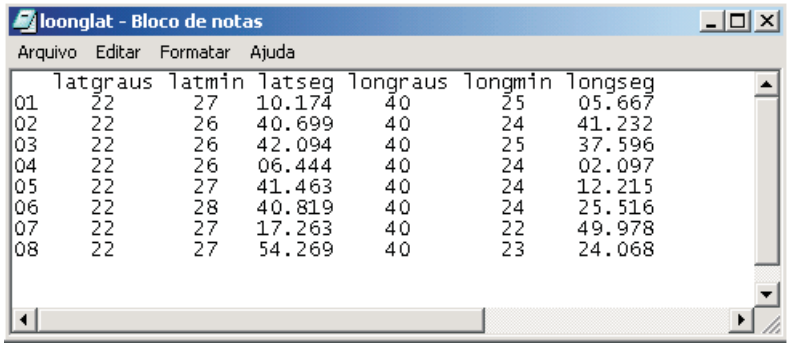


Figura 4.2: Dados utilizados no formato adequado.

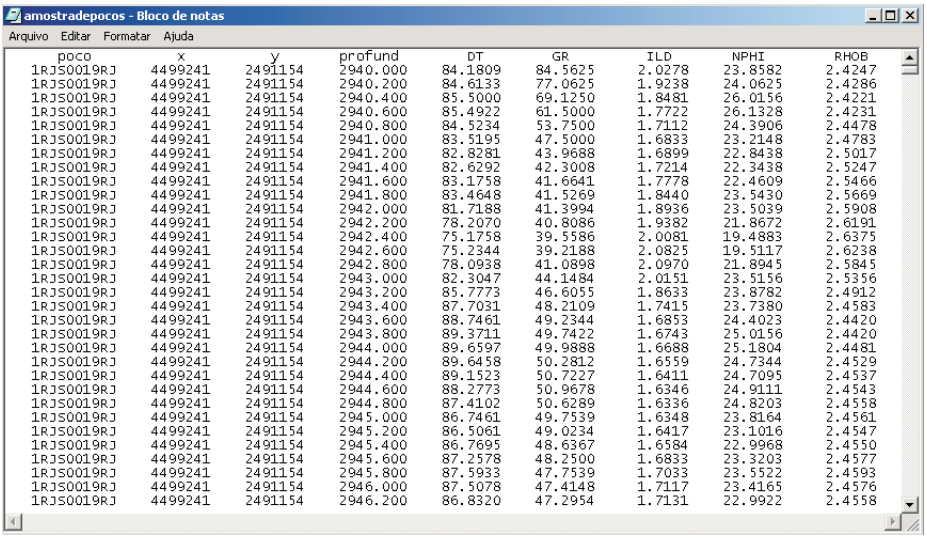
Em seguida, os dados de latitude e longitude foram convertidos (Veja a rotina no Programa 1 do Apêncice A).

4.3 Formatação dos Dados

A etapa subseqüente foi a junção dos dados de localização dos poços obtidos no Programa 1 com os demais dados, em um único arquivo, denominado de *amostradepocos.txt*. Seguindo a mesma formatação dos dados da Figura 4.2, como pode ser observado na Figura 4.3, os dados de cada poço foram reduzidos proporcionalmente, de maneira que o total de registros que apresentassem informações a respeito desses oito poços fosse reduzida de 7825 para um número aproximadamente igual a mil (número de dados razoável na realização de análises através do *software* R, para o tipo de análise desejado). Após a redução, o número de registros encontrado foi 1294. Veja a rotina do programa para a redução de dados no Programa 2 do Apêndice A.

4.4 Gerando Resultados

Após a formatação e redução dos dados, foram obtidos os variogramas experimentais de cada variável, considerando como coordenadas as posições x, y e *profund*. A rotina referente a esta parte da pesquisa encontra-se no Programa 3, no Apêndice A.



amostradepocos - Bloco de notas

Arquivo Editar Formatar Ajuda

poco	x	y	profund	DT	GR	ILD	NPHI	RHOB
1RJS0019RJ	4499241	2491154	2940.000	84.1809	84.5625	2.0278	23.8582	2.4247
1RJS0019RJ	4499241	2491154	2940.200	84.6133	77.0625	1.9238	24.0625	2.4286
1RJS0019RJ	4499241	2491154	2940.400	85.5000	69.1250	1.8481	26.0156	2.4221
1RJS0019RJ	4499241	2491154	2940.600	85.4922	61.5000	1.7722	26.1328	2.4231
1RJS0019RJ	4499241	2491154	2940.800	84.5234	53.7500	1.7112	24.3906	2.4478
1RJS0019RJ	4499241	2491154	2941.000	83.5195	47.5000	1.6833	23.2148	2.4783
1RJS0019RJ	4499241	2491154	2941.200	82.8281	43.9688	1.6899	22.8438	2.5017
1RJS0019RJ	4499241	2491154	2941.400	82.6292	42.3008	1.7214	22.3438	2.5247
1RJS0019RJ	4499241	2491154	2941.600	83.1758	41.6641	1.7778	22.4609	2.5466
1RJS0019RJ	4499241	2491154	2941.800	83.4648	41.5269	1.8440	23.5430	2.5669
1RJS0019RJ	4499241	2491154	2942.000	81.7188	41.3994	1.8936	23.5039	2.5908
1RJS0019RJ	4499241	2491154	2942.200	78.2070	40.8086	1.9382	21.8672	2.6191
1RJS0019RJ	4499241	2491154	2942.400	75.1758	39.5586	2.0081	19.4883	2.6375
1RJS0019RJ	4499241	2491154	2942.600	75.2344	39.2188	2.0825	19.5117	2.6238
1RJS0019RJ	4499241	2491154	2942.800	78.0938	41.0898	2.0970	21.8945	2.5845
1RJS0019RJ	4499241	2491154	2943.000	82.3047	44.1484	2.0151	23.5156	2.5356
1RJS0019RJ	4499241	2491154	2943.200	85.7773	46.6055	1.8633	23.8782	2.4912
1RJS0019RJ	4499241	2491154	2943.400	87.7031	48.2109	1.7415	23.7380	2.4583
1RJS0019RJ	4499241	2491154	2943.600	88.7461	49.2344	1.6853	24.4023	2.4420
1RJS0019RJ	4499241	2491154	2943.800	89.3711	49.7422	1.6743	25.0156	2.4420
1RJS0019RJ	4499241	2491154	2944.000	89.6597	49.9888	1.6688	25.1804	2.4481
1RJS0019RJ	4499241	2491154	2944.200	89.6458	50.2812	1.6559	24.7344	2.4529
1RJS0019RJ	4499241	2491154	2944.400	89.1523	50.7227	1.6411	24.7095	2.4537
1RJS0019RJ	4499241	2491154	2944.600	88.2773	50.9678	1.6346	24.9111	2.4543
1RJS0019RJ	4499241	2491154	2944.800	87.4102	50.6289	1.6336	24.8203	2.4558
1RJS0019RJ	4499241	2491154	2945.000	86.7461	49.7539	1.6348	23.8164	2.4561
1RJS0019RJ	4499241	2491154	2945.200	86.5061	49.0234	1.6417	23.1016	2.4547
1RJS0019RJ	4499241	2491154	2945.400	86.7695	48.6367	1.6584	22.9968	2.4550
1RJS0019RJ	4499241	2491154	2945.600	87.2578	48.2500	1.6833	23.3203	2.4577
1RJS0019RJ	4499241	2491154	2945.800	87.5933	47.7539	1.7033	23.5522	2.4593
1RJS0019RJ	4499241	2491154	2946.000	87.5078	47.4148	1.7117	23.4165	2.4576
1RJS0019RJ	4499241	2491154	2946.200	86.8320	47.2954	1.7131	22.9922	2.4558

Figura 4.3: Junção dos dados de localização com os dados das variáveis.

A consideração dos valores do *efeito pepita*, do *alcance* e do *patamar*, no variograma experimental dos dados, e a posterior associação deste a um variograma teórico com estas características, constituem uma etapa fundamental na análise geoestatística de um conjunto de dados. É através do variograma teórico encontrado que é feita a krigagem, através da qual os dados não amostrados podem ser inferidos. A precisão desta inferência está diretamente ligada à variabilidade dos dados e à correlação existente entre os mesmos, observados nos valores do *patamar* e do *alcance*, encontrados no variograma empírico.

Como os dados utilizados nas análises foram capturados a cada $20\,cm$ de profundidade, e selecionados arbitrariamente na fase de redução de dados, temos que a menor distância entre eles é de $20\,cm$. Além disso, cálculos realizados revelaram que a maior das distâncias entre eles é $5195.738\,cm$. Por serem valores muito discrepantes, o ajuste dos variogramas empíricos obtidos aos respectivos variogramas teóricos não foi realizado, já que os dados mais próximos estavam sendo considerados como tendo a mesma localização. Uma alternativa para resolver este problema seria a utilização de dados cuja distância mínima entre eles seja necessariamente bem maior que $20\,cm$.

Por consequência, a validação do modelo e a krigagem também não foram feitas.

Os resultados obtidos foram os sequintes:

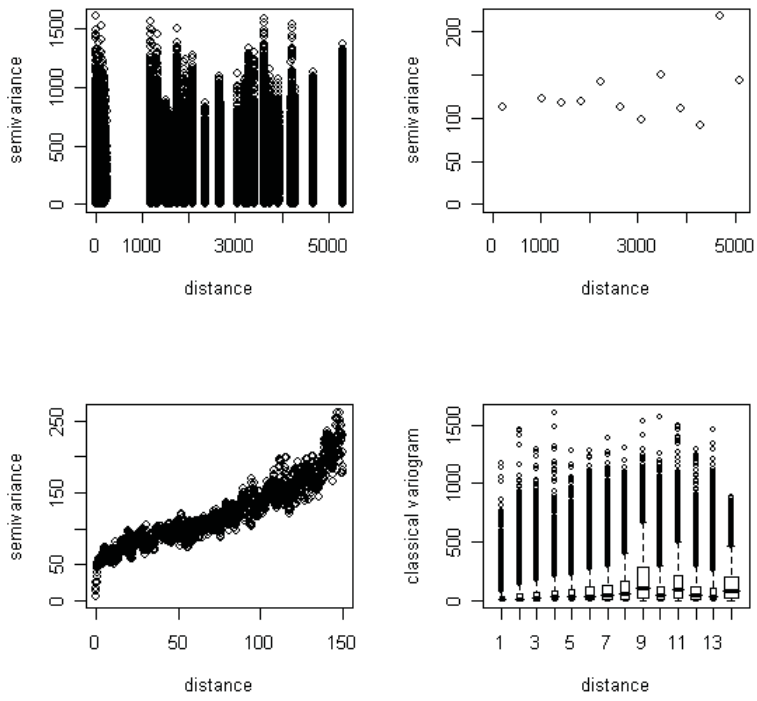


Figura 4.4: Gráficos gerados para a variável DT.

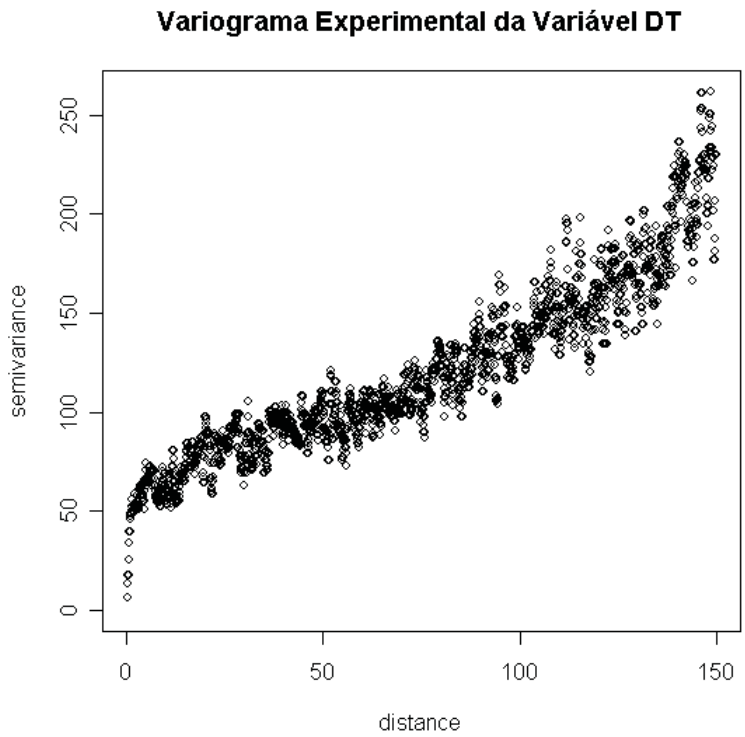


Figura 4.5: Variograma da variável DT.

O comportamento dos dados do variograma experimental da variável DT (Figuras 4.4 e 4.5) não pode ser comparado a um modelo de variograma teórico conhecido. Nota-se que, para distâncias maiores que um metro, o variograma experimental adquire características de um modelo linear. Porém, isto não acontece para valores de distância menores que um metro. Uma alternativa seria considerar alguns tipos de modelos de variogramas teóricos ou combinações destes, para testar qual deles resulta em um menor erro na krigagem.

O variograma experimental da variável GR (Figuras 4.6 e 4.7) sugere o modelo esférico, com *efeito pepita* $= 0$, *patamar* $= 200$ e *alcance* $= 12\,m$. Isto significa que existe uma correlação entre os valores de GR, desde que os pontos analisados estejam situados a uma distância igual ou inferior a 12 metros. Apesar deste resultado ser válido para distâncias menores ou iguais a 12 metros, tanto verticalmente quanto horizontalmente, devemos restringi-lo apenas às distâncias verticais (ou seja, no mesmo poço), tendo em vista que a menor distância entre os poços analisados é de, aproximadamente, 831 metros. De fato, as grandes distâncias entre os poços utilizados na análise são justificadas pelos poços serem quase todos de extensão da jazida devendo, assim, delimitar toda a área que deve ser posta em produção. O único poço analisado que não é de extensão é o poço pioneiro.

A distância entre os poços perfurados em uma jazida depende de vários fatores, inclusive da viscosidadse do fluido existente na rocha reservatório. Quanto mais viscoso o fluido, menor tende a ser a distância entre os poços através dos quais ele chegará à superfície. No entanto, existe um limite para a proximidade entre os poços, dados os altos custos dos processos de perfuração e produção. Assim sendo, mesmo que estivessem sendo considerados outros tipos de poços, como por exemplo os poços de desenvolvimento do reservatório, seria improvável haver pelo menos dois poços perfurados a uma distância igual ou inferior a 12 metros um do outro.

O valor do *patamar* indica uma grande variabilidade entre os dados, se comparados aos demais valores de *patamar* para os variogramas experimentais das demais variáveis.

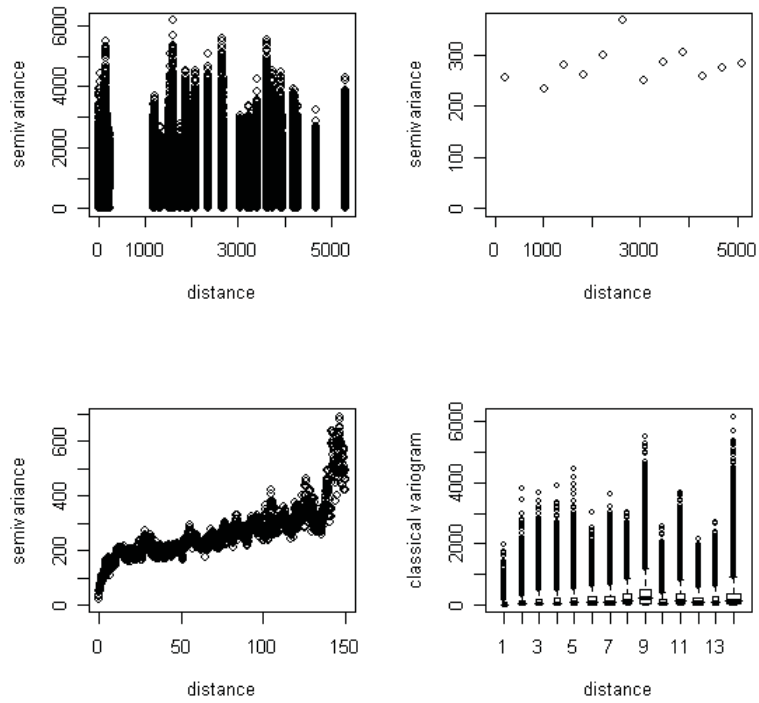


Figura 4.6: Gráficos gerados para a variável GR.

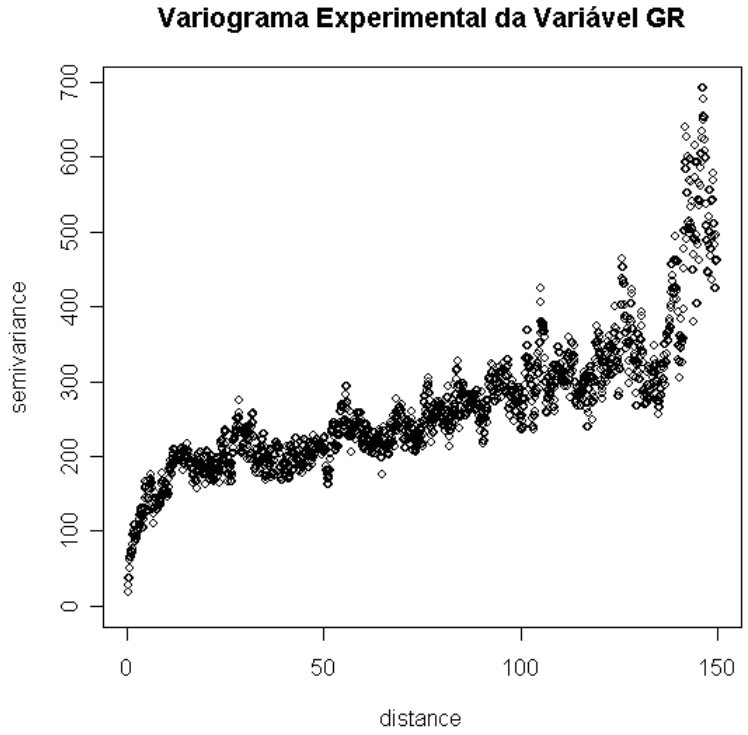


Figura 4.7: Variograma da variável GR.

Por se tratar de uma variável cujos valores são altamente discrepantes e por não possuir nenhum valor nulo, os dados da variável ILD sofreram uma transformação prévia para que os resultados não fossem mascarados. Ao invés de analisarmos a variável ILD, utilizamos o logaritmo de ILD. O variograma empírico desta variável (Figuras 4.8 e 4.9), assim como o da variável GR, revelou apresentar modelo esférico, com *efeito pepita* $= 0$, *patamar* $= 0.61$ e *alcance* $= 25\,m$, ambos na escala logarítmica. Pelas mesmas razões citadas nas análises de GR, o valor do anlcance deve refletir uma correlação apenas vertical de 25 metros. Isto significa que acima deste valor os dados perdem a correlação.

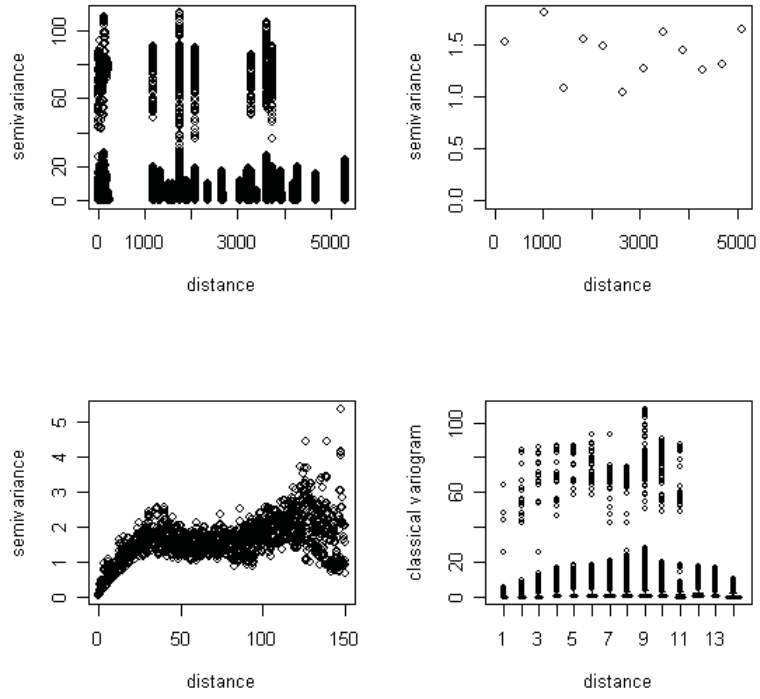


Figura 4.8: Gráficos gerados para o logaritmo da variável ILD.

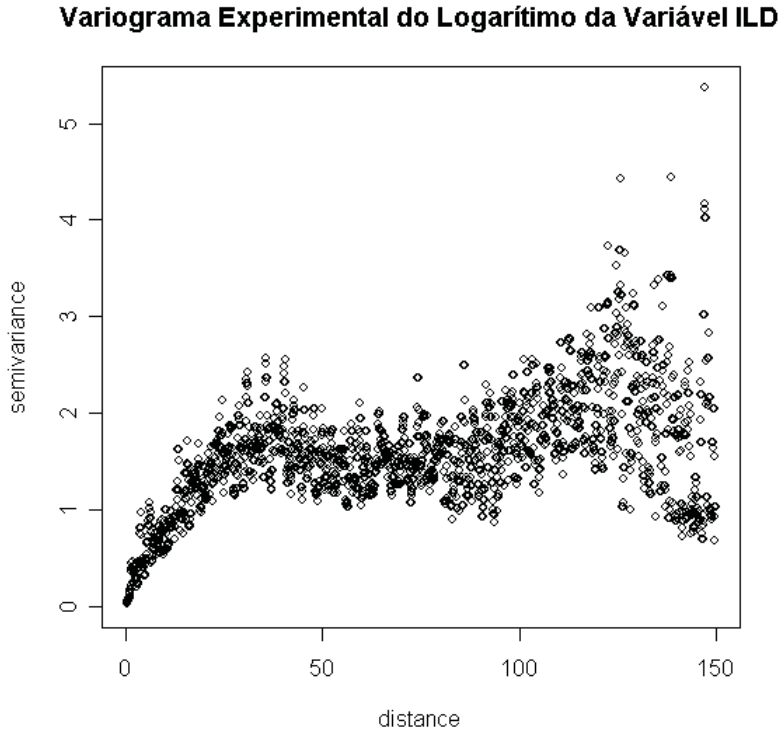


Figura 4.9: Variograma do logaritmo da variável ILD.

Convém enfatizar que a variável GR é a variável que possui a segunda maior variabilidade entre os dados. A variável cujos valores são mais discrepantes é a variável ILD. O valor encontrado para o *patamar* do variograma experimental de ILD está na escala logarítmica.

O variograma experimental dos dados da variável NPHI (Figuras 4.10 e 4.11) possui características semelhantes ao variograma dos dados da variável DT.

A análise variográfica da variável RHOB (Figuras 4.12 e 4.13) nos leva a um modelo esférico, com *efeito pepita* $= 0$, *patamar* $= 0.0182$ e *alcance* $= 18\,m$. Podemos observar que, dentre as variáveis analisadas, esta foi a que resultou em um menor *patamar*. Isto significa que os dados de RHOB não são tão discrepantes quanto os dados das demais variáveis. Além disso, o *alcance* de $18\,m$ indica que existe uma correlação entre os dados quando entre estes não houver uma distância superior a 18 metros.

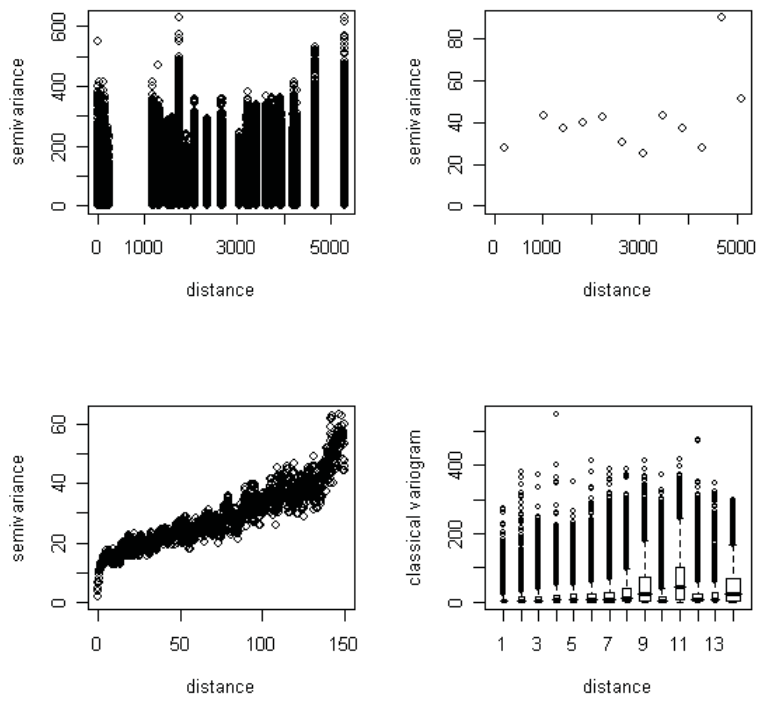


Figura 4.10: Gráficos gerados para a variável NPHI.

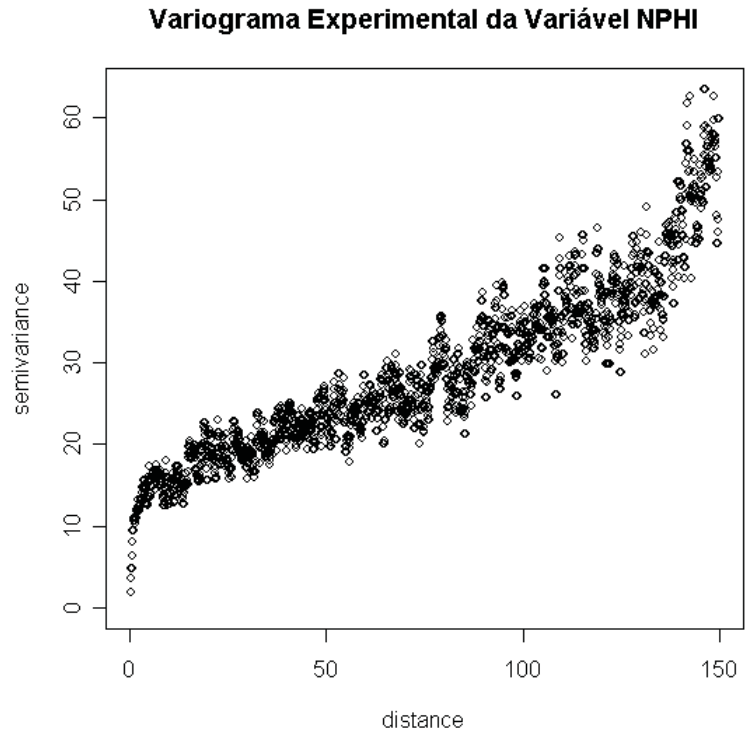


Figura 4.11: Variograma da variável NPHI.

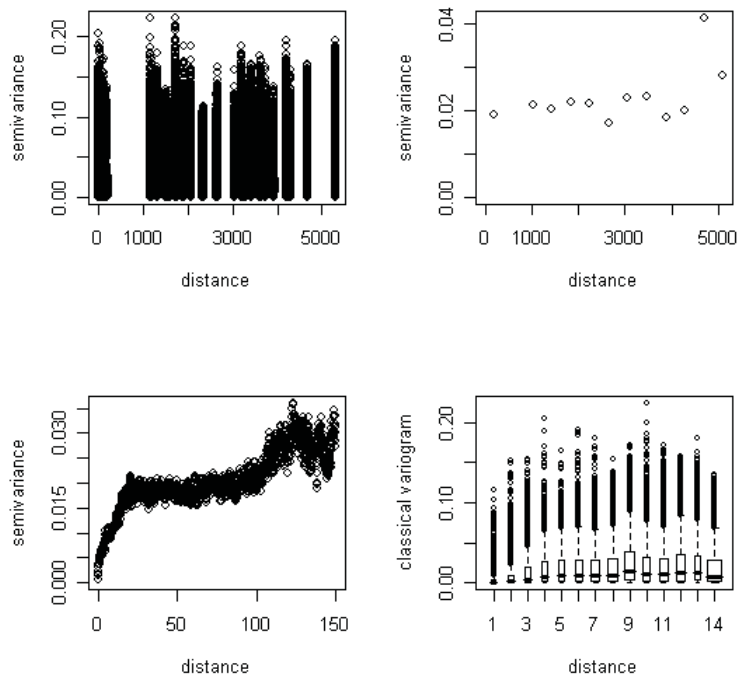


Figura 4.12: Gráficos gerados para a variável RHOB.

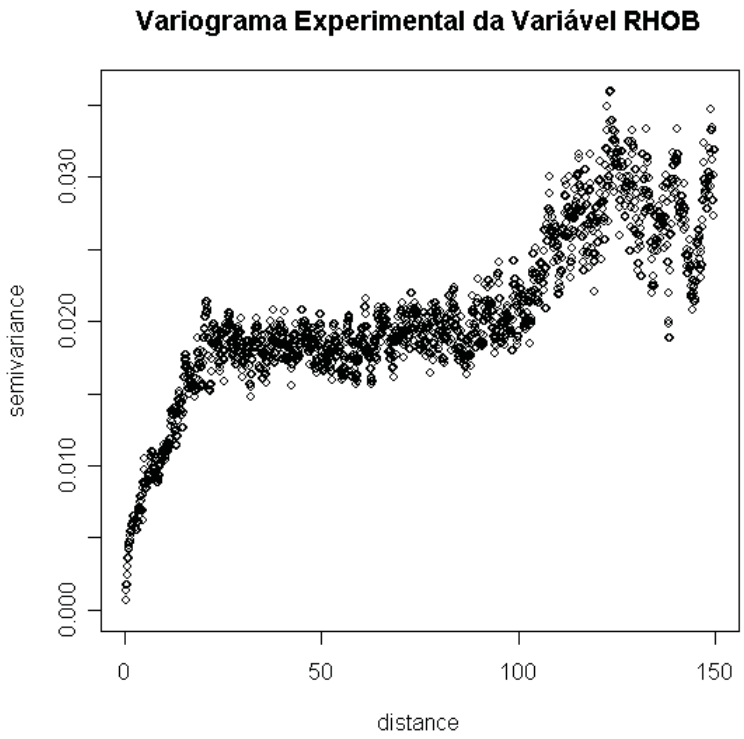


Figura 4.13: Variograma da variável RHOB.

Em geral, pode-se observar que quanto maior o valor do *patamar*, maior a discrepância entre os dados, podendo ocasionar erros maiores na estimação dos valores não observados. Outra informação importante é que quanto maior o valor do *alcance*, mais distantes podem estar os pontos correlacionados entre si.

Certamente, as etapas não realizadas no desenvolvimento deste trabalho, isto é, a validação do modelo e a krigagem, não devem ser deixadas de lado ou esquecidas, pois apesar dos custos operacionais na realização de análises como estas serem muito altos, estes são certamente superados pelos benefícios alcançados.

Por levar em consideração a localização espacial dos dados e o fenômeno da anisotropia, mencionados no Capítulo 3, a análise geoestatística produz resultados bastante plausíveis, com uma margem de erro menor que a margem de erro fornecida pelas análises que utilizam as técnicas estatísticas.

Conclusão

A utilização da simulação como ferramenta de análise permite projetar diversos cenários, suprindo limitações dos dados e reduzindo custos na amostragem, objetivando reproduzir comportamentos de características que foram identificadas na análise estrutural [1]. Em particular a simulação estocástica, uma vez que poucos fenômenos geológicos são compreendidos o suficiente de modo a permitir uma abordagem determinística para realizar estimativas [17]. A Geoestatística produz numerosos resultados plausíveis, os quais requerem um certo esforço por parte do pesquisador. Entretanto, os benefícios superam o tempo e custos adicionais. Há incertezas nos pontos não amostrados e a abordagem geoestatística baseia-se em modelos probabilísticos que reconhecem e incorporam estas incertezas inevitáveis. Um melhor conhecimento do campo petrolífero torna a sua exploração uma atividade de menor risco.

Apêndice A

Rotinas dos Programas

A.1 Programa 1: Transformação das Coordenadas dos Poços

```
#---------------------------------------------------------------------#
# Arquivo de Dados (longlat.txt):

   latgraus latmin latseg longraus longmin longseg
01    22      27   10.174    40       25    05.667  # Dados de 1RJS0019RJ
02    22      26   40.699    40       24    41.232  # Dados de 3NA001ARJS
03    22      26   42.094    40       25    37.596  # Dados de 3NA0002RJS
04    22      26   06.444    40       24    02.097  # Dados de 3NA0003RJS
05    22      27   41.463    40       24    12.215  # Dados de 3NA0004RJS
06    22      28   40.819    40       24    25.516  # Dados de 3NA005ARJS
07    22      27   17.263    40       22    49.978  # Dados de 3NA017ARJS
08    22      27   54.269    40       23    24.068  # Dados de 3NA021BRJS
#---------------------------------------------------------------------#

# Definindo o perímetro da Terra na linha do Equador (usado no cálculo
# da longitude)
# 2*pi*R1; R1=6378000 metros
c1<-2*3.14159265359*6378000

# Definindo o perímetro da Terra no Meridiano de Greenwhich (usado
# no cálculo da latitude). Note que o raio da Terra no sentido
```

```
# Norte-Sul é menor que o raio da terra no sentido Leste-Oeste.
# 2*pi*R2; R2=6357000 metros
c2<-2*3.14159265359*6357000

# Definindo um grau com relação à longitude:
graulong<-c1/360

# Definindo um grau com relação à longitude:
graulat<-c2/360

# Entrando com os dados de latitude e longitude dos poços
mdados<-read.table("longlat.txt",header=TRUE)
attach(mdados)

# Obtendo o valor de n(número de poços) (No nosso caso, n=8)
n<-nrow(mdados)

for (i in 1:n)
+ {
+ # Transformando os dados da forma "grau-min-seg" para a forma "total
+ # de seg"
+ lattseg<-(3600*mdados[,1])+(60*mdados[,2])+mdados[,3]
+ longttseg<-(3600*mdados[,4])+(60*mdados[,5])+mdados[,6]
+
+ # Relacionando quantos graus estão associados aos dados de entrada
+ # de longitude
+ grauslong<-longttseg/3600
+
+ # Relacionando quantos graus estão associados aos dados de entrada
+ # de latitude
+ grauslat<-lattseg/3600
+ }

# Encontrando as coordenadas dos poços(em metros)
x<-graulong*grauslong
y<-graulat*grauslat
loc<-cbind(x,y)

# Imprimindo as localizações dos poços
```

```
loc
        x          y
[1,] 4499241    2491154     # poço 1RJS0019RJ
[2,] 4498486    2490246     # poço 3NA001ARJS
[3,] 4500229    2490289     # poço 3NA0002RJS
[4,] 4497276    2489190     # poço 3NA0003RJS
[5,] 4497589    2492118     # poço 3NA0004RJS
[6,] 4498000    2493948     # poço 3NA005ARJS
[7,] 4495046    2491373     # poço 3NA017ARJS
[8,] 4496127    2492542     # poço 3NA021BRJS
```

A.2 Programa 2: Organização dos Dados

```
#-------------------------------------------------------------------------#
# Leitura dos dados:

oitopocos<-read.table("amostradepocos.txt", header=TRUE)
attach(oitopocos)
#-------------------------------------------------------------------------#

# Selecionando uma amostra aleatória das medidas obtidas nas diversas
# profundidades em cada poço:

# Amostra de tamanho 149, do poço 1RJS0019RJ:
amostra_1<-sample(1:900, size=149, replace=FALSE, prob=NULL)
# Amostra de tamanho 207, do poço 3NA001ARJS:
amostra_2<-sample(901:2150,size=207,replace=FALSE, prob=NULL)
# Amostra de tamanho 186, do poço 3NA0002RJS:
amostra_3<-sample(2151:3275, size=186, replace=FALSE, prob=NULL)
# Amostra de tamanho 91, do poço 3NA0003RJS:
amostra_4<-sample(3276:3825, size=91, replace=FALSE, prob=NULL)
# Amostra de tamanho 165, do poço 3NA0004RJS:
amostra_5<-sample(3826:4825, size=165, replace=FALSE, prob=NULL)
# Amostra de tamanho 165, do poço 3NA005ARJS:
amostra_6<-sample(4826:5825, size=165, replace=FALSE, prob=NULL)
# Amostra de tamanho 157, do poço 3NA017ARJS:
amostra_7<-sample(5826:6775, size=157, replace=FALSE, prob=NULL)
```

```
# Amostra de tamanho 174, do poço 3NA021BRJS:
amostra_8<-sample(6776:7825, size=174, replace=FALSE, prob=NULL)
#-----------------------------------------------------------------------#

# Ordenando cada amostra, e associando cada linha selecionada à informação
# do arquivo de dados:

A<-oitopocos[sort(amostra_1),]
B<-oitopocos[sort(amostra_2),]
C<-oitopocos[sort(amostra_3),]
D<-oitopocos[sort(amostra_4),]
E<-oitopocos[sort(amostra_5),]
F<-oitopocos[sort(amostra_6),]
G<-oitopocos[sort(amostra_7),]
H<-oitopocos[sort(amostra_8),]
#-----------------------------------------------------------------------#

# Armazenando o novo arquivo de dados:
contemplados<-rbind(A,B,C,D,E,F,G,H)

# Exportando os dados para um arquivo de texto do Bloco de Notas:
write.table(contemplados, file = "contemplados.txt", append = FALSE,
            quote = TRUE, sep = "    ", eol = "\n", na = "NA",
            dec = ".", row.names = FALSE, col.names = TRUE,
            qmethod = c("escape", "double"))
```

A.3 Programa 3: Variogramas Experimentais para Cada Variável

```
#-----------------------------------------------------------------------#
# Leitura dos dados:
contemplados  <- read.table("contemplados.txt", header=T)
contemplados  <- data.frame(contemplados)

#-----------------------------------------------------------------------#
```

```
# Vaiável DT

# Transformando os dados para o formato geodata, utilizado pelo pacote
# geoR, do software R:

geoDT.data <- as.geodata(contemplados, coords.col=c(2,3,4), data.col=c(5))
summary(geoDT.data)

matriz.de.distancias <- dist(geoDT.data$coords)
probs = c(0, 1, 2, 3, 4, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.5, 25)/100
pontos.u <- quantile(matriz.de.distancias, probs = probs, na.rm =
+ FALSE, type = 7)

win.graph()
par(mfrow=c(2,2))
plot( variog(geoDT.data, option="cloud"))
plot( variog(geoDT.data, option="bin", bin.cloud=TRUE))
plot( variog(geoDT.data, option="smooth", uvec=pontos.u, max.dist=150))
plot( variog(geoDT.data, option="bin", uvec=pontos.u, bin.cloud=TRUE),
+ bin.cloud=T)

#-------------------------------------------------------------------------#
# Variável GR

geoGR.data <- as.geodata(contemplados, coords.col=c(2,3,4), data.col=c(6))
summary(geoGR.data)

matriz.de.distanciasGR <- dist(geoGR.data$coords)
pontos.uGR <- quantile(matriz.de.distanciasGR, probs = probs, na.rm =
+ FALSE, type = 7)

win.graph()
par(mfrow=c(2,2))
plot( variog(geoGR.data, option="cloud"))
plot( variog(geoGR.data, option="bin", bin.cloud=TRUE))
plot( variog(geoGR.data, option="smooth", uvec=pontos.u, max.dist=150))
plot( variog(geoGR.data, option="bin", uvec=pontos.uGR, bin.cloud=TRUE),
+ bin.cloud=T)
```

```
#-----------------------------------------------------------------------#
# Variável ILD

geologILD.data <- as.geodata(contemplados, coords.col=c(2,3,4), data.col=c(7))
summary(geologILD.data)

matriz.de.distanciaslogILD <- dist(geologILD.data$coords)
pontos.uILD <- quantile(matriz.de.distanciaslogILD, probs = probs, na.rm =
+ FALSE, type = 7)

win.graph()
par(mfrow=c(2,2))
plot( variog(geologILD.data, option="cloud"))
plot( variog(geologILD.data, option="bin", bin.cloud=TRUE))
plot( variog(geologILD.data, option="smooth", uvec=pontos.u, max.dist=150))
plot( variog(geologILD.data, option="bin", uvec=pontos.uILD, bin.cloud=TRUE),
+ bin.cloud=T)

#-----------------------------------------------------------------------#
# Variável NPHI

geoNPHI.data <- as.geodata(contemplados, coords.col=c(2,3,4), data.col=c(8))
summary(geoNPHI.data)

matriz.de.distanciasNPHI <- dist(geoNPHI.data$coords)
pontos.uNPHI <- quantile(matriz.de.distanciasNPHI, probs = probs, na.rm =
+ FALSE, type = 7)

win.graph()
par(mfrow=c(2,2))
plot( variog(geoNPHI.data, option="cloud"))
plot( variog(geoNPHI.data, option="bin", bin.cloud=TRUE))
plot( variog(geoNPHI.data, option="smooth", uvec=pontos.u, max.dist=150))
plot( variog(geoNPHI.data, option="bin", uvec=pontos.uNPHI, bin.cloud=TRUE),
+ bin.cloud=T)
```

```
#-----------------------------------------------------------------------#
# Variável RHOB

geoRHOB.data <- as.geodata(contemplados, coords.col=c(2,3,4), data.col=c(9))
summary(geoRHOB.data)

matriz.de.distanciasRHOB <- dist(geoRHOB.data$coords)
pontos.uRHOB <- quantile(matriz.de.distanciasRHOB, probs = probs, na.rm =
+ FALSE, type = 7)

win.graph()
par(mfrow=c(2,2))
plot( variog(geoRHOB.data, option="cloud"))
plot( variog(geoRHOB.data, option="bin", bin.cloud=TRUE))
plot( variog(geoRHOB.data, option="smooth", uvec=pontos.u, max.dist=150))
plot( variog(geoRHOB.data, option="bin", uvec=pontos.uRHOB, bin.cloud=TRUE),
+ bin.cloud=T)
```

Bibliografia

[1] Braga, L. P. V., *Geoestatística e Aplicações*. 9º Simpósio Nacional de Probabilidade e Estatística, São Paulo, (1990).

[2] Burrough, P. A., *Principles of geographical information systems for land resources assessment*. Clarendon Press, Oxford, (1987).

[3] Bussab, W. O., Miazaqui, E. S. e Andrade, D. F., *Introdução à Análise de Agrupamentos*. Associação Brasileira de Estatística, 9° Simpósio Nacional de Probabilidade e Estatística. São Paulo, (1990).

[4] Chambers, R. L., Yarus, J. M. e Hird, K. B., *Petroleum geostatistics for nongeostaticians*. Geologic Column of The Leading Edge, (2000).

[5] Costa, A. P. A., *Desenvolvimento de um Simulador Térmico para Recuperação de Petróleos Viscosos Via Aquecimento Eletromagnético*. Dissertação (Mestrado), Engenharia Química, UFRN, (1998).

[6] David, M., *Geostatistical ore reserve estimation*. Elsevier Scientific, New York, (1977).

[7] Davis, J. C., *Statistics and Data Analysis in Geology*, (1973).

[8] Fernandes, C. E. M., *Fundamentos de Prospecção Geofísica*. Ed. Interciência ,Rio de Janeiro - RJ, (1984).

[9] Fisher, R. A., *The Statistical Utilization of Multiple Measurements*. Annals of Eugenics, (1938).

[10] Guerra, P. A. G., *Geoestatística Operacional*. Departamento Nacional da Produção Mineral, Brasília, (1988).

[11] Isaaks, E. H. e Srivastava, R. M., *An Introduction to Applied Geostatistics*. Oxford University Press, New York, (1989).

[12] Johnson, R. A. e Wichern, D. W., *Applied Multivariate Statistical Analysis*. 5^{nd} Edition, Prentice-Hall, New York, (2002).

[13] Journel, A. G., *Fundamentals of geostatistics in five lessons*. Standford Center for Reservoir Forecasting Applied Earth Sciences Department, California, (1988).

[14] Landim, P. M. B., *Análise Estatística de Dados Geológicos*, Fundação Editora da UNESP, São Paulo, (1998).

[15] Lourenço, A. e Matias, R. P. *Estatística Multivariada*. Instituto Superior de Engenharia do Porto, (2000).

[16] Rocha, A. C. B., *Agrupamento de Poços Petrolíferos do Campo Escola de Namorado por Análise Estatística de Perfis*. Monografia de Graduação, PRH(25)-ANP/MME/MCT, UFCG, Campina Grande - PB, (2002).

[17] Santos, C. G. P., Mato, L. F. e Clennell, B., *Modelagem Estocástica Aplicada à Caracterização do Reservatório do Campo de Namorado*. Anais do 2° Congresso Brasileiro de P&D em Petróleo e Gás, Rio de Janeiro - RJ, (2003).

[18] Santos, M. B., *Modelagem Estocástica Baseada em Objetos de Reservatórios Turbidíticos Canalizados*. Dissertação de Mestrado, Instituto de Geociências - UNICAMP, (2002).

[19] Swan, A. R. H., Sandilands, M., *Introduction to Geological Data Analysis*. Blackwell Science Ltd, (1995).

[20] Thomas, J. E. e outros, *Fundamentos de Engenharia de Petróleo*. Rio de Janeiro, Editora Interciência, PETROBRAS, (2001).

[21] http://laqqa.iqm.unicamp.br/PCA2.htm

[22] http://www.dpi.inpe.br/spring/usuario/variograma.htm

[23] http://www.eps.ufsc.br/teses96/camargo/anexo/apendice2.htm

[24] http://www.hydrolog.com.br/internas/perfilagem/perfil_1.shtm

[25] http://www.sbgf.org.br/geofisica/geofisica.html