Geoestatística

Introdução aos princípios

David Alvarenga Drumond

David Alvarenga Drumond

Geoestatística - Introdução aos princípios

1 edição

Belo horizonte
22/09/2017

D249i Drumond, David
Geoestatística - Introdução aos princípios /
David A. Drumond. - Belo Horizonte : O Autor, 2017.
117 p.

ISBN: 978-85-923922-0-8

1. Engenharia de Minas . 2. Geologia. 3. Geoestatística.
I. Título.

CDD: 550

Aos meus pais pelo eterno carinho e apoio.

Sempre que te perguntarem se podes fazer um trabalho,
respondas que sim e te ponhas em seguida a aprender como se faz.
F. Roosevelt

0.1 Introdução

Este livro é apenas uma breve iniciação aos conceitos matemáticos e definições da geoestatística, uma área dos fenômenos estocásticos que lida com variáveis aleatórias espaciais. O objetivo desta obra é auxiliar no primeiro contato com esta disciplina, tentando demonstrar da maneira mais simples o funcionamento da teoria, primeiramente concebida por Matheron, e depois formalizada por uma gama de autores, que tornou a geoestatística uma ferramenta poderosa na avaliação de vários processos de engenharia, ciências e até mesmo medicina. No entanto, este é um livro criado principalmente para estudantes de graduação dos cursos voltados para as áreas de mineração. O primeiro capítulo envolverá uma série de conceitos sobre planejamento mineral e geologia, de forma a incluir o leitor no assunto. O segundo capítulo trata da formalização da teoria das variáveis aleatórias regionalizadas. Em seguida são introduzidos conhecimentos de estatística básica envolvendo o tratamento univariado e bivariado. Técnicas de desagrupamento são então iniciadas, demonstrando ao leitor a importância da prática de amostragem. Por conseguinte as funções de continuidade espacial e variografia são conceituadas, para então serem introduzidas as técnicas de estimativa espacial também conhecida como krigagem. Por final, o livro apresenta as curvas de teor e tonelgagem que representam um resumo final da avaliação do depósito mineral. É importante lembrar que este livro é apenas introdutório e não representa um sumário de todas as técnicas geoestatísticas possíveis. Na verdade, esta obra se limita apenas a um entendimento básico do planejamento mineral e muitas outras técnicas foram desenvolvidas para lidar com problemas multivariados ou simulação de depósitos minerais. No entanto, mesmo sendo um livro simples, ainda representa um alicerce na construção do conhecimento geoestatístico pelo leitor e muito do que ainda se pratica no meio da mineração é ainda observado com estas metodologias simples.

Conteúdo

1. Estimativa de depósitos minerais

Este capítulo inicial é uma breve introdução aos conceitos de planejamento mineral e avaliação de recursos e reservas. O objetivo com este texto é iniciar o leitor nos jargões mais comuns da mineração facilitando seu entendimento na utilização da geoestatística em seu meio de trabalho. A estimativa de depósitos minerais é sem dúvida o campo mais importante do uso da geoestatística no setor mineral, lidando com as questões de relevância no planejamento e que impactam diretamente em toda a cadeia de processo. É importante observar que provavelmente todas as variáveis lidadas na mineração são estocásticas, ou seja, apresentam natural variabilidade na sua medição. No entanto, a estimativa de depósitos minerais significa uma das incertezas que mais possui capacidade de impacto no planejamento de uma mina. Um estudo profundo sobre geoestatística é quase uma necessidade de qualquer planejador.

1.0.1 Introdução

Os investimentos necessários para iniciar uma mineração tendem a ser da ordem de grandeza de milhões de reais. De forma a obter um investimento rentável, o material produzido pela mina e posteriormente beneficiado deve ser potencialmente adequado em quantidade e qualidade necessária para justificar a decisão do investimento.

A mineração e o processamento mineral devem operar de forma a planejar um lucro aceitável. Certamente todas a tecnologia e as decisões financeiras são tomadas visando a viabilidade técnica, ambiental e social do commodite. Logo a estimativa dos teores deve ser conhecida com certo intervalo de segurança, de forma a propiciar uma garantia sobre o lucro inferido. Dessa forma a geoestatística prevalece como o conjunto de metodologias que melhor adéqua as estimativas do depósito mineral. Apenas garantindo condições de segurança aos investidores externos, por meio de amostragens corretas e cálculos confiáveis é possível atrair investimentos financeiros para o empreendimento mineral.

Isso é de fato real em todos as condições de operação de uma mina, seja em depósitos extensos e disseminados, ao qual os teores são apenas ligeiramente acima das condições de lucratividade, e para materiais preciosos onde somente existe um percentual pequeno mineralizado, que pode ser explotado com certo lucro.

Os lucros na mineração são altamente nivelados pelo preço de mercado e pelo teor do material minerado. Uma pequena diferença entre os valores planejados e realizados ou uma pequena diferença no preço do metal associado tem um grande impacto na lucratividade da mina.

Para permanecer competitiva, as companhias de mineração devem otimizar sua produtividade em cada operação unitária. Há várias formas de se conseguir este objetivo. Movimentando ou processando mais toneladas de material com o menos equipamento é uma das alternativas, seguido de um melhor controle das operações ou comprando equipamentos mais eficientes. Todas essas formas de operação estão associadas com um custo e com um potencial de retorno do investimento. A redução de custos poussui um papel importante na determinação da lucratividade, pois fatores como a flutuação de preços ou as incertezas geológicas são parâmetros muitas vezes incontroláveis. Outra forma de agregar valor ao commodite é potencializando o conteúdo metálico do produto, utilizando rotas de beneficiamento mais eficientes, o que torna o desenvolvimento tecnológico uma grande provedor de recursos da economia mineral.

Os três pilares para o controle das incertezas na mineração e otimização das operações de mineração são: estimativa do minério, planejamento mineral e controle dos teores. O nosso interesse neste livro é focar na estimativa do minério para uma mina. A geoestatística é o conjunto de técnicas atual que melhor averigua as incertezas geológicas e constrói uma visão acerca das possibilidades máximas de execução da cadeia mineral.

1.0.2 Estimativa dos depósitos mienrais

Jazidas minerais são consideradas uma quantificação formal da ocorrência de materiais naturais, que são estimados por uma variedade de procedimentos tanto epíricos como teóricos. Estas são consideradas a base do estudo de viabilidade econômica e são classificadas como reservas e recursos. Considera-se um recurso mineral quando apenas estima-se possíveis quantidades de material capazes de gerar um lucro para o empreendimento. Já uma reserva mineral é uma quantidade limitada que possui um nível de confiança suficiente para garantir lucratividade, disponibilidade técnica, jurídica, ambiental e social.

Os recursos e reservas minerais são determinados a partir de amostras de rochas que determinam um volume de rocha mineralizado de uma ordem de grandeza muito maior do que o volume da amostra. Esses erros de estimativa são vistos como erros de extensão ao considerar a ampliação dos valores da amostra para o todo. Com o objetivo de caracterizar um depósito mineral, este é dividido em um conjunto de blocos, ao qual o valor estimado de uma propriedade de cada bloco está relacionado com os dados mais proximais. Logo a jazida mineral pode ser vista como um quebra-cabeça com tamanhos individuais, com localização e propriedades estabelecidas. Existem, na verdade, diversas formas de apresentação dos valores de uma propriedade do depósito mieneral. A Figura (1.1) demonstra um exemplo. Algumas delas como polígonos de influência e triangulação de Delunay representam antigas formas de estimativa de um depósito mineral, mas que, no entanto, ainda têm hoje papel na geoestatística. A opção de uma configuração de blocos geométricos definidos é necessária considerando o conceito de suporte, caracterizado posteriormente, e pelas condições de operacionalização da mina.

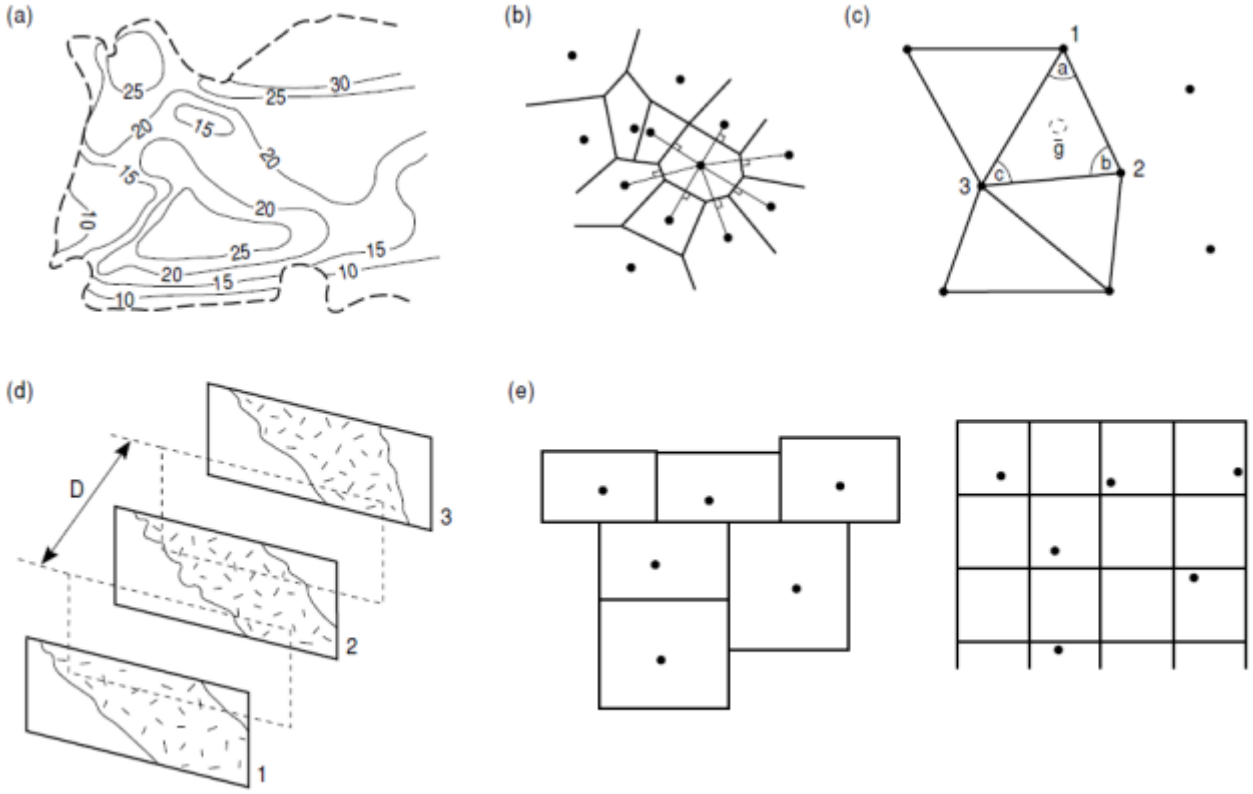


Figura 1.1: Figura demonstrando diversas apresentações de uma propriedade do depósito mineral. a) isolinhas b) Polígonos de influência c) triangulação d) seção paralelas e) blocos irregulares f) blocos regulares

Definimos então um importante conceito na engenharia mineral, denominado de unidade seletiva de lavra.

R A unidade seletiva de lavra é o menor volume a ser retirado de um depósito mineral, previamente estimado, segundo as condições técnicas e econômicas viáveis à sua extração.

Logo entendemos que uma mina é antes de apenas uma escavação na terra, uma concepção abstrata de propriedades de interesse desconhecidas de um volume de um corpo geológico.

A quantificação de recursos e reservas minerais exige um certo grau de confiabilidade (subjetivo ou estatístico) apropriado para os dados disponíveis durante a estimativa. Volumes, massa, teores e quantidades de metal ou minerais atributos são geralmente quantificados.

1.0.3 Alguns conceitos iniciais sobre jazidas minerais

Na engenharia de mina, tal como na estimativa de depósitos minerais e planejamento mineral, há uma série de jargões técnicos utilizados. É importante compreendê-los de forma a planificar o entendimento sobre o assunto tratado. Alguns conceitos, no entanto, podem até mesmo apresentar sentido ambíguo, como o conceito de continuidade espacial, que dentro da geologia representa uma medida da estrtutura física de um corpo mineral e na geoestatística representa um grau de similaridade entre variáveis aleatórias. Isso muitas vezes ocorre devido ao uso de palavras

semelhantes ou idênticas em setores diferenciados da mineração. Apesar da atividade mineradora ser um processo contínuo de transformação da matéria, ela é segmentada em diversos subsetores interdependentes.

Minério

Minério é uma rocha ou mineral que é extraído por trabalhos de mineração com o intuito (mesmo que algumas vezes não alcançado) de obter vantagens para a comunidade. Dentre estas vantagens pode-se citar o benefício econômico, a produção exclusiva para construção civil, a obtenção de recursos energéticos para o Estado ou na pesquisa científica.

O termo minério é aplicado a rochas em três formas mais comuns. 1) Como uma descrição econômica, relacionada com o controle de qualidade das reservas minerais e com o seu conteúdo 2) Como um commodite vendido à parte do seu conteúdo, tal como em pedras de construção 3) Ou como um material fragmentado oriundo de uma mina.

O primeiro tópico é o de maior importância e implica na distinção de minério (material minerado com lucro) e estéril (que não contêm valor suficiente para se obter lucro). A definição de minério e estéril é uma função dependente do tempo que relaciona diversos fatores tais como preço, tecnologia, regime de taxação, condições ambientais, sociais, etc.

Em geral minas são colocadas em atividade com o entendimento que será possível um retorno necessário ao seu investimento. As circunstâncias ditarão a concepção do preço e consequentemente daquilo caracterizado como minério e estéril. Para reduzir os efeitos da incerteza do tempo um a mineração geralmente trabalha antecipadamente com três formas de planejamento, a curto, médio e longo prazo.

R Minério é uma rocha ou mineral que é extraída com a finalidade de proporcionar um benefício para a sociedade. Diferentemente do mineral, o minério é uma concepção econômica, técnica e temporal e não é necessariamente ligada à entidade física que a compõe.

Teor de corte

O conceito de teor de corte (ou cutoff) é definido como aquele em que o valor do conteúdo metálico ou mineral, em um certo volume de rocha, começa a atender as especificações econômicas da mina. Os teores de corte são usados para distinguir blocos de minério e estéril em vários estágios da evolução da estimativa da jazida mineral (exploração, desenvolvimento e produção). Minério/Estéril são baseados nos valores estimados. Em alguns casos os erros de estimativa podem levar a uma classificação errada do material. A figura (1.2) demonstra quatro regiões definidas pelos erros de estimativa. O material classificado como minério pode ser estéril ou de fato minério, enquanto o material classifcado com estéril pode ser minério ou de fato estéril.

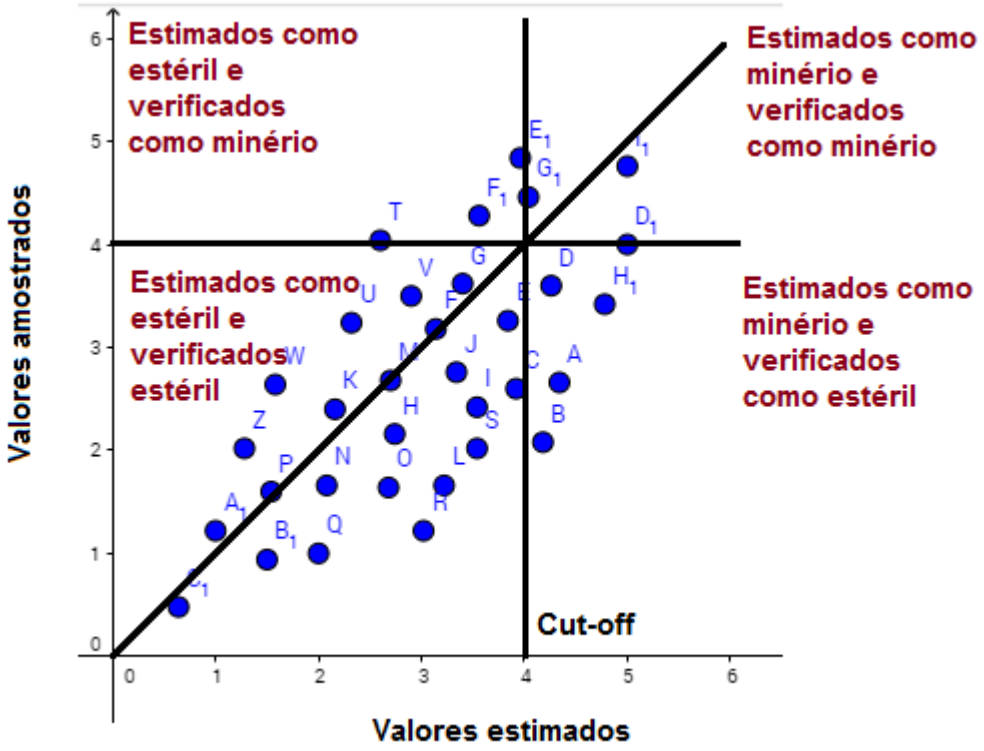


Figura 1.2: Figura demonstrando o gráfico de aderência entre os valores estimados e verificados posteriormente por amostragem. Quatro regiões determinadas pelo valor de teor de corte

Com o aumento do teor de corte, a massa de minério tende a diminuir e o teor médio da jazida acima do teor de corte tende a aumentar. Geralmente com o aumento da razão estéril e minério (unidades de estéril a ser removido por unidade de minério retirado) também aumenta com o teor de corte. Geralmente apenas um pequeno espectro de teores de corte são considerados no processo de simulação e seleção de um cenário particular de mineração. Alternativas recentes de planejamento mineral tendem a incorporar o máximo possível do espectro de variabilidade das funções de transferência do depósito.

O conceito de teor de corte também está ligado com a conectividade dos minérios durante o estágio de produção. Com o aumento do teor de corte o volume de minério tende a se dividir de forma a criar volumes separados na jazida.

A estimativa do teor de corte é na verdade um problema econômico complexo, e está fora do escopo deste livro, no entanto podemos apresentá-la de forma simplificada. O custo operacional de um minério por tonelada beneficiada pode ser dado pela equação (1.1)

$$OC = FC + (SR + 1)MC \tag{1.1}$$

Em que FC é o custo fixo, SR é a relação estéril minério e MC é o custo de mineração por tonelada movimentada.

O fluxo de caixa de um mina pode ser dado pela relação (1.2)

$$CF = Receita - Custosoperacionais = (g.F.P - OC)T \tag{1.2}$$

Onde g é o teor médio do minério, F é a recuperação metalúrgica e P é o preço por tonelada

beneficiada e T a tonelagem de material beneficiado. Logo podemos encontrar o teor de corte quando o Fluxo de caixa é igual a zero (1.3)

$$g = \frac{OC}{F.P} \tag{1.3}$$

R Teor de corte pode ser considerado como a menor proporção de um dado elemento de interesse na rocha capaz de classificá-la como rentável (minério) ou não-rentável (estéril)

Continuidade

Continuidade pode ser definido como o grau de conectividade no espaço. Na avaliação de depósitos este termo é utilizado ambiguamente como uma ocorrência física ou natural geológica, tal como uma estrutura, um litotipo ou uma mineralização ou como o grau de correlação espacial entre variáveis aleatórias. Um capítulo neste livro é dedicado somente a caracterização da continuidade estatística de variáveis aleatórias.

Diluição

A diluição é o resultado da mistura do minério com estéril durante o processo de produção, geralmente causando um aumento do volume de material retirado e decréscimo do valor médio do teor segundo as expectativas. Pode-se dividir a diluição em duas categorias: interna (material de baixo teor envolvido de material com alto teor) ou externa (material de alto teor envolvido com material de baixo teor).

Recursos e reservas minerais

Jazidas geralmente são consideradas em termos de recursos ou reservas minerais. As definições geralmente variam segundo uma jurisdição para outra, no entanto há um grande esforço para tornar os conceitos internacionalizados. Na falta de consenso internacional, há uma tendência tando industrial como técnica de adotar o código australiano ou JORC(Joint Ore Reserves Committee).

Um recurso é uma ocorrência mineral quantificada alicerçado nos dados geológicos e no teor de corte simplesmente. Diversas são as formas de se classificar os recursos minerais, sejam elas baseadas simplesmente na geometria das amostras, na variância de krigagem ou em simulações geoestatísticas. Na maioria dos códigos não existe metodologias prescritas primordialmente. Espera-se apenas que se utilize um critério que garanta confiabilidade nas estimativas. Os recursos são subdivididos em medido, indicado e inferido segundo o grau de confiabilidade das medidas. Essas classificações podem ser feitas segundo uma distância geométrica do centro da amostra, de intervalos para o valor da variância de krigagem em uma região ou para um nível de confiabilidade para a distribuição de uma parcela do depósito mineral. Espera-se que o valor medido tenha maior confiabilidade, o indicado menor e o inferido pouca certeza.

Recursos podem ser transformados em reservas minerais caso haja um estudo de viabilidade do depósito mineral. Este incorpora uma variedade de aspectos relacionados com a lucratividade do bem mineral e estão ligados com as questões ambientais , técnicas, sociais, jurídicas e financeiras do empreendimento. As reservas são dividas em provável e provados.

A classificação das regiões do depósito são variáveis ao longo do desenvolvimento do projeto e das fases de desenvolvimento da mina, devido ao acréscimo de informação. Um recurso pode se tornar uma reserva mineral e vice-versa. A figura (1.3) demonstra a transição das diversas classificações.

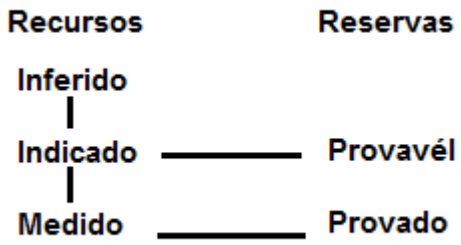


Figura 1.3: Figura demonstrado a classificação de jazidas em recursos e reservas. Linhas indicando a transição entre as classificações

Unidade Seletiva de mineração

Uma unidade seletiva de mineração é o menor bloco ao qual se define um material como minério ou estéril, ou seja, o menor volume de rocha no planejamento capaz de ser retirado e decidido sua finalidade, seja para a usina de beneficiamento, seja para a pilha de estéril. O tamanho da unidade seletiva de mineração depende do método de lavra utilizado e da escala de produção da operação. Para objetivos de planejamento um depósito mineral pode ser considerado um arranjo de blocos definidos por esta unidade, cada um com seu valor associado de teor e outros parâmetros como densidade, massa, volume, saturação de água, etc.

Na decisão do tamanho de bloco duas condições são impostas para a determinação do seu volume. A primeira são as condições operacionais da mineração, tal como altura do talude, volume a ser retirado com o caminhão, dimensão da malha de desmonte, espaçamento durante a produção e etc. A segunda é o tamanho mínimo para uma estimativa confiável. Esta geralmente é calculada como 1/4 da distância média entre as amostras. A sub-blocagem, ou seja, a divisão dos blocos para atender os requisitos operacionais não deve ser feita durante a fase de estimativa, mas apenas sobre os blocos estimados, como forma de adequar o processo à operação.

Precisão e Exatidão

Exatidão pode ser exemplificado como a proximidade de uma estimativa com a realidade, enquanto precisão é a medida da dispersão entorno de uma estimativa. Uma estimativa pode ser exata, mas não precisa, ou precisa, mas não exata. A figura (1.4) demonstra os conceitos de exatidão e precisão a partir de figuras de alvos. O centro do alvo é o valor verdadeiro que pretende-se alcançar com os disparos. Disparos entorno do centro são considerados exatos, enquanto disparos próximos aos outros são considerados precisos.

Figura 1.4: Figura demonstrando os conceitos de exatidão e precisão. O centro do alvo é o valor verdadeiro que pretende-se alcançar com os disparos. Disparos entorno do centro são considerados exatos. Disparos próximos aos outros são considerados precisos

Na estimativa de depósitos minerais é natural que os nossos disparos não sejam precisos, mas é mais que importante que sejam exatos. Não é admitido obter valores médios com erros deslocados da média global. Isto também é chamado de viés ou deriva.

Há vários tipos de erros potenciais na estimativa de reservas minerais incluindo:

- Erro de amostragem
- Erros de análise química.
- Erros de densidade (É comum em muitos casos considerar a densidade do material constante ao longo do depósito)
- Erros da geologia, durante as fases de determinação da continuidade espacial e geometria do depósito mineral.
- Na escolha do método de lavra adotado que pode não atender as questões de seletividade do minério e estéril de forma ótima.
- A diluição do minério com a encaixante.
- Erro humano (inserção de valores errados no banco de dados, de casas decimais, et.)
- Fraude (salgamento de amostras, substituições de amostras, dados não representativos, etc.)

1.1 Softwares de Mineração

Durante a fase de avaliação das jazidas minerais o computador exerce função essencial como ferramenta de estudo. Uma quantidade substancial de softwares estão disponíveis em meio comercial e alguns aplicativos livres também existem. Softwares comerciais são mais onerosos, mas possuem suporte técnico e manutenção de seus sistemas. Apresentam código fechado ao público externo pertencente geralmente ao proprietários. Softwares gratuitos geralmente são disponibilizados por universidades, possuem código aberto ao público e podem ser facilmente obtidos via Internet.

Uma das bibliotecas gratuitas mais importantes é sem dúvida o GSLIB (Geostatistical Software Library) e apresenta além dos executáveis do programa seus algoritmos, escritos em fortran 90 e disponibilizados no site. Os programas são administrados pelo doutor Clayton Deutsch e Emmanuel Schnetzler. Mais informações sobre o pacote de softwares pode ser encontrado no site www.gslib.com ou no guia de uso [4]

O uso dos softwares de mineração geralmente requerem que os arquivos de dados sejam organizados eficientemente em formatos pré-estabelecidos, gerados pelas campanhas de exploração.

Essa compilação dos dados é trabalhosa e necessita de uma validação primordial, tornando o trabalho de preparação dos dados às vezes muito mais demorado que várias implementações dos programas.

Entre as aplicações mais comuns encontradas em softwares de mineração relacionadas com a estimativa de depósitos temos:

- Uma grande variedade de procedimentos de avaliação dos dados (estatísticas, gráficos, etc.:)
- Determinação da qualidade dos dados e dos protocolos de amostragem
- Modelagem tridimensional e visualização de formas geológicas complexas e distribuição das amostras.
- Preparação de seções planas e verticais
- Gráficos de contorno tanto do teor como de outras variáveis
- Caracterização da continuidade espacial (Variogramas automáticos, mapas de variograma, variogramas experimentais e modelagem)
- Modelagem de blocos do depósito
- Metodologias de cálculos de recurso e reservas
- Avaliações dos efeitos de vários métodos de mineração
- Determinação da viabilidade econômica de depósitos

Alguns destes softwares podem ainda incluir ferramentas de planejamento de mina, tal como otimização de cava, sequenciamento, desenho de cava, etc.

A grande desvantagem da maioria dos softwares, principalmente dos pagos é algumas questões referentes ao seu funcionamento que não são explicadas pelos manuais. Alguns modelos matemáticos, são de fato, "escondidos"dentro das rotinas dos programas. Isso dificulta a tomada de decisão dos operadores em alguns casos e pode até ser prejudicial em algumas formas.

1.2 Exercícios

Exercise 1.1 Realize um "brainstorm"e pense todas as possibilidades que podem sofrer uma mina que possam tornar um minério em um estéril. Por exemplo, a descoberta de uma outra jazida de uma empresa concorrente mais próximo do mercado consumidor pode aumentar o preço do minério e tornar parte do recurso inutilizável por um tempo. E quais seriam os fatores que fazem um estéril se tornar minério? ■

Exercise 1.2 Pretende-se determinar se uma unidade seletiva de lavra é um minério ou estéril. O custo fixo de extração do material é 5 um/ton. O custo de mineração por tonelada movimentada é 2 um/ton. A relação estéril/minério é 3/2. A Recuperação metalúrgica é de 95% e o preço do minério é de 100 um/ton. O teor do elemento útil do bloco é 2%. R. Teor de corte 10% -> estéril. ■

Exercise 1.3 Os dados da tabela seguinte demonstram um conjunto de valores estimados e dados reais obtidos. Determine:

a) O viés das estimativas. (Diferença entre a média dos valores estimados e a dos reais)

b) Considere o cut-off como 2g/ton. Determine: A proporção dos valores estimados como minério que realmente são minério. A proporção dos valores estimados como estéreis que realmente são estéreis.

Estimados	Real
2.05	2.0
2.03	2.02
1.01	1.32
2.31	3.45
3.02	1.02
2.76	2.19
3.08	4.01
3.74	3.67
1.02	1.43
1.00	1.01
2.03	1.05

2. Variáveis aleatórias regionalizadas

Este primeiro capítulo trata de uma introdução à geoestatística, abordando sua importância, necessidade e os primeiros aspectos relacionados com a teoria das variáveis aleatórias regionalizadas. Iniciamos o conceito de variável e de momentos estatísticos, que são de grande importância para a compreensão dos capítulos seguintes. Maiores informações podem ser encontradas nas obras de Matheron [10] ou nos livros base de [8] e [6]

2.1 Introdução

A geoestatística é um conjunto de técnicas que utiliza a teoria das variáveis aleatórias regionalizadas como uma ferramenta para a descrição, estimativa e avaliação de fenômenos espaciais. É portanto, um grupo de modelos matemáticos, e não possuí características de um modelo físico ou realista, mas apenas uma descrição e avaliação probabilística acerca do fenômeno estudado.

Seu início remota aos anos de 1950, quando D.G.Krige concluiu, na África do Sul, que não poderia estimar de forma adequada o conteúdo de ouro em blocos mineralizados se porventura não considerasse a geometria, posicionamento e volume das amostras. Esse conceito, posteriormente definido pelos geoestatísticos como suporte, foi a base para o desenvolvimento do engenheiro francês George Matheron na criação alicerces da teoria.

Até meados daquela época os métodos de avaliação de recursos e reservas minerais utilizavam-se apenas da posição geométrica e disposição das amostras. Os chamados métodos clássicos foram abolidos pelas geoestatística porque não ofereciam determinadas vantagens como garantir um modelo de controle estrutural, determinar formas de se avaliar a qualidade da estimativa ou garantir pesos diferentes para amostras agrupadas nas estimativas.

É notório que, devido à sua origem, a geoestatística possua ampla utilização no setor mineral, mas também é aplicada na exploração petrolífera, engenharia ambiental e civil, além de demais campos das ciências médicas, biológicas e geografia. Na verdade qualquer fenômeno disposto no espaço, com uma componente aleatória, pode ser tratável no âmbito da teoria das variáveis aleatórias regionalizadas. A utilização das técnicas se faz ainda mais preponderante quando o atributo estudado é escasso de amostras e informações, o que pode ser constatado principalmente na engenharia de mina e de petróleo, ao qual a amostragem é onerosa e esparsa.

No setor mineral a geoestatística tem representação em diversas áreas, desde a pesquisa com grande peso na avaliação de recursos e reservas, na amostragens do depósito, nas usinas de beneficiamento e até mesmo na avaliação de operações unitárias tais como desmonte e homogeneização de pilhas de minério. É impossível se pensar na formação moderna de um engenheiro de minas sem conhecimentos básicos sobre a geoestatística, devido à sua amplitude como ferramenta nas tomadas de decisão técnica.

O caso mais comum de sua utilização talvez seja a criação de um modelo de blocos para um elemento metálico de interesse. Valores são inferidos em locais onde não há amostragens e posteriormente é utilizado um planejamento de cava e sua operacionalização. Todo o procedimento de decisão é realizado a partir desta estimativa, o que torna as incertezas geológicas um dos parâmetros que mais podem afetar um projeto mineral. Consegue-se então definir um recurso e este será a base para investimentos e captação financeira de uma empresa. A responsabilidade para determinar estimativas confiáveis geralmente é atribuída a um competent person (pessoa competente).

Um aspecto comum dos problemas geoestatísticos é a amostragem. Seja na mineração, na exploração de petróleo ou em outra área, as estimativas realizadas sobre o fenômeno estão sempre baseada em um conjunto de informações retiradas do todo. As informações obtidas podem ser advindas de testemunhos de sondagem, poços, trincheiras, pó de perfuratriz ou dados geofísicos. É importante considerar que cada amostra advindo de fontes de pesquisa diferentes possuem qualidades estatísticas díspares e por isso devem ser tratadas separadamente. Teores medidos por testemunhos de sondagem, por exemplo, devem ser tratados como uma amostra diferente das obtidas por pó de perfuratriz, mesmo que o elemento medido seja o mesmo.

O conjunto de técnicas geoestatísticas que utilizam informações secundárias para beneficiar as estimativas é chamado de geoestatística multi-variada.

Na geoestatística assume-se que as amostras possuam valor determinístico caracterizado por um único valor médio em um suporte determinado. No entanto, é de se esperar que a amostragem gere erros associados, o que significa que os modelos à posteriori também incorporem estes valores. A amostragem possui influência direta nos métodos geoestatísticos e na avaliação de depósitos minerais, e é responsabilidade do avaliador entender dos protocolos e criticá-los. Erros nos valores médios de amostragem alcançam em alguns casos ordens de 10-20 por cento e podem ser decisivos nas metodologias de avaliação.

É importante ao leitor entender que a geoestatística é simplesmente um conjunto de modelos. Por ser um modelo ela pretende aproximar a realidade a partir de uma descrição matemática simplificada. Modelos mais complexos conseguem reproduzir melhor a realidade do atributo a ser modelado. Modelos mais simples não conseguem reproduzir toda a complexidade da realidade, mas são mais rápidos e fáceis de serem colocados em prática. Por mais esforço que se empenhe em um modelo matemático ele nunca será a realidade.

Dentre os modelos da geoestatística podemos citar:

1. Geoestatística univariada e multivariada: Definem-se os diversos modelos que podem utilizar apenas uma variável ou a utilização de múltiplas variáveis correlacionadas.
2. Geoestatística linear e não-linear: Definem-se os diversos modelos que podem utilizar variáveis que possam ser modeladas por modelos lineares e as que necessitam de uma transformação à priori para serem utilizadas.

A geoestatística multivariada, ao utilizar uma maior quantidade de informação sobre a variável de interesse, consegue traduzir com melhor perfeição as características do fenômeno descrito. Em muitos casos como em engenharia de petróleo, por exemplo, nem mesmo a informação direta de poços é abundante. A informação secundária tal como uma sísmica de reflexão pode ser a informação chave para a definição de estruturas geológicas e na modelagem dos litotipos. Em outros casos, uma variável secundária pobre ou mal amostrada pode não interferir na definição de

um modelo mais robusto, tornando a estimativa apenas uma tarefa mais trabalhosa, sendo mais adequado utilizar um modelo univariado mais simples.

Nem todas as variáveis tratadas são somáticas. Em alguns casos como o teor, sabemos que a quantidade de um elemento metálico em um bloco de uma mina, somado com a quantidade de elemento metálico de outro bloco, resulta no total de metal dos blocos. Isso não pode ser realizado com algumas variáveis como por exemplo, a condutibilidade hidráulica de uma rocha. A média aritmética da condutibilidade de dois blocos não é em suma o valor médio da condutibilidade dos mesmos. Se uma rocha é impermeável e outra conduz líquido, não poderíamos esperar que em média as duas conduzissem líquido. Neste caso devemos utilizar a chamada geoestatística não-linear.

A geoestatística não-linear é a base para os modelos de simulação geoestatística. Neste caso estamos interessados em não apenas encontrar os valores mais prováveis de um bloco de minério, mas todos os valores equiprováveis de bloco, para que tenhamos noção do espectro de incerteza relacionado com nossa estimativa.

O objetivo nesta primeira introdução é demonstrar o que é a geoestatística, qual é a sua importância dentro da mineração, quais problemas ela tem como premissa resolver e como ela funciona dentro de um entendimento como modelo. As seções posteriores definirão conceitos chaves para o entendimento das variáveis aleatórias regionalizadas e os primeiros aspectos matemáticos envolvidos neste livro.

2.2 Variável aleatória regionalizada

Define-se uma variável aleatória como sendo uma função em que se associa cada elemento do conjunto universo à um valor real, e a este uma dada probabilidade. Matematicamente expressamos a variável aleatória como $Z : \Omega \rightarrow \Re$. O lançamento de dados, por exemplo, pode assumir valores $1,2,3,4,5,6$ com valores de probabilidade iguais a $1/6$. Por via de regra definimos uma variável aleatória com uma letra maiúscula Z, e uma realização como o valor de 1 no dado, com uma letra minúscula z=1.

Uma variável aleatória regionalizada é uma variável aleatória associada à um suporte x no espaço. Entende-se que x possuí posição, geometria e volume. Podemos identificar x como as coordenadas cartesianas, por exemplo, tal que $x = [x,y,z]$, ou como coordenadas cilíndricas $x = [r,\theta]$. O valor de x também poderia ser a coordenada do centro de massa de um bloco ou painel de uma mina. O que define a importância de um suporte como pontual, bidimensional e tridimensional é o problema tratado. Os testemunhos de sondagem de um corpo de minério, quando vistos na dimensão do depósito podem ser considerados quase que unidimensionais ao longo de sua extensão.

É importante lembrar que a variável aleatória é dispare para cada posição geométrica e volume considerado no espaço. A variável aleatória no suporte $x = [0,0,0]$ é diferente da variável aleatória $x = [0,0,1]$ podendo assumir valores distintos e probabilidades diferentes.

As variáveis aleatórias podem ser contínuas, quando seus valores podem estar contidas dentro dos intervalos reais, podem ser discretas quando assumirem valores inteiros. Diversas são as variáveis aleatórias disponíveis na mineração, tais como recuperação metalúrgica, teor de um elemento metálico, condutibilidade hidráulica, saturação de água, etc.

Uma peculiaridade da variável aleatória regionalizada é que ela assume valor determinístico nos locais amostrados. Em outras palavras, no ponto amostrado, o único valor possível da variável aleatória regionalizada é o valor da amostra.

Quando a variável aleatória regionalizada é definida para todo o domínio do espaço então temos uma função aleatória $Z(x)$. Na geoestatística não estamos geralmente interessados em definir a função aleatória e a distribuição de probabilidades para qualquer suporte no domínio considerado.

Para nós é mais importante conhecer o grau de dependência de uma variável aleatória com outra em sua proximidade.

Duas variáveis aleatórias são consideradas independentes quando a sua covariância é igual a zero. Nos fenômenos geológicos tais como em muitos daqueles descritos pela natureza, os fenômenos mais proximais tendem a ser mais dependentes. Isso significa que em um derramamento de magma gerador de um depósito vulcanogênico, o material mais próxima da borda do vulcão tende a ser mais semelhante que o mais distal.

Determinando assim uma lei de conectividade ou correlação da função aleatória $Z(x)$ podemos tentar prever uma realização de uma variável aleatória $Z(x_i)$ em um local desconhecido. Este valor é indeterminado, mas no entando, podemos garantir que uma dada realização tenha uma maior probabilidade de existir, mesmo que isso não ocorra! Por isso devemos entender, antes de tudo, que estimativas de um depósito mineral são medidas estocásticas e não analíticas, logo não podem ser consideradas exatas. Tudo o que se pode fazer é garantir que as possibilidades estejam mais possivelmente a seu favor, mas nunca previsões exatas da realidade.

2.3 Hipótese de estacionaridade

A estacionaridade é uma das hipóteses da metodologia geoestatística que permitem a simplificação de algumas de suas formulações. A chamada estacionaridade intrínseca é aquela que é inerente da teoria e propõe que o valor esperado de um incremento de duas variáveis aleatórias regionalizadas $Z(x_i) - Z(x_j)$ depende exclusivamente de um vetor h associado. Esta também é chamada de invariância por translação. Em outras palavras dizemos que a correlação entre as variáveis aleatórias regionalizadas não muda com a posição considerada, e depende somente do vetor associado. A utilização de um modelo vetorial possui sérias implicações na geoestatística. Toda a informação utilizada para a inferência em locais desconhecidos é realizada entre variações médias de pontos já conhecidos, o que implica que a qualidade do modelo está muito mais ligado com o posicionamento geométrico das amostras do que necessariamente com os valores obtidos pela amostragem.

Outra hipótese de estacionaridade também considerada é a hipótese de estacionaridade de segunda ordem. Esta implica que o valor médio da função aleatória é constante ao longo de todo o domínio. $m = E(Z(x))\forall x \in D$, sendo D o domínio considerado. Da mesma forma podemos dizer para a sua variância. No entanto, há casos não-estacionários em que o valor médio da função aleatória tende a se propagar segundo uma direção. Quando a variância aumenta proporcionalmente com a média chamamos este efeito de "efeito proporcional".

Observe a figura (2.1). A imagem é um exemplo de como podemos interpretar a questão da estacionaridade de segunda ordem de uma variável aleatória regionalizada. O perfil topográfico de um terreno apresenta uma parte estacionária, ao qual os valores de cota tendem a flutuar em um patamar e uma parte não estacionária, ao qual os valores de cota tendem a decrescer continuamente.

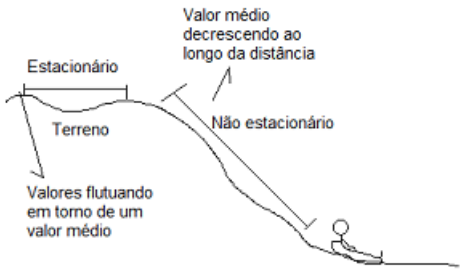


Figura 2.1: Exemplo estacionaridade com um perfil topográfico. Região estacionária é plana, enquanto a região não-estacionária apresenta caimento constante.

2.4 Decomposição da função aleatória

Toda função aleatória $Z(x)$ pode ser dividida em dois componentes: um valor de tendência e um de resíduo. Tal como uma combinação de variáveis aleatórias regionalizadas a função aleatória também é uma variável aleatória, possuindo um valor médio e uma distribuição de probabilidades para cada suporte x considerado. Como definido na seção anterior sob a hipótese de estacionaridade de segunda ordem temos um valor médio constante da função aleatória por todo o domínio. Denotaremos m como sendo o valor desta tendência (trend) da função aleatória. Retirado o valor médio desta função aleatória temos então o resíduo, dependente do suporte x considerado. Logo definimos que:

$$Z(x) = Y(x) + m(x) \tag{2.1}$$

Em que $m(x)$ é o valor médio da função aleatória dependente da posição no espaço x e $Y(x)$ é o resíduo. Em muitos casos na mineração temos uma estrutura de controle no depósito mineral que cria condições de formação dado um viés ao longo do espaço. Uma falha na rocha ou um enriquecimento supergênico pode fazer com que determinado elemento se concentre ou disperse ao longo de uma direção.

2.5 Momentos estatísticos

Momentos estatísticos são funções que caracterizam o comportamento de uma variável aleatória. Como estas assumem valores diferentes de acordo com sua probabilidade associada, podemos tentar definir seu valor mais provável, o grau de dispersão destes valores entre outros parâmetros.

Duas propriedades mais comuns de uma variável aleatória são seus momentos de inércia de primeira e segunda ordem. Estes são definidos pelo valor esperado. A definição de valor esperado de uma variável aleatória pode ser representado pela equação (2.2)

$$E(Z) = \int_{-\infty}^{+\infty} f(z)\, z dz \tag{2.2}$$

Onde z é uma realização da variável aleatória $f(z)$ é a sua função de densidade de probabilidade associada. Em outros termos a esperança matemática é nada mais que uma média ponderada da

variável aleatória pela sua probabilidade, ou o valor mais provável de uma variável aleatória. Na geoestatística estamos interessados em estimar este parâmetro a partir de amostras, o que torna conveniente atribuir o valor esperado em sua forma discreta tal como demonstrado na equação (2.3)

$$E(Z) = \sum_{i=-\infty}^{+\infty} p(z_i) z_i \tag{2.3}$$

Em que $p(z_i)$ é a probabilidade de ocorrência da realização z_i. A esperança matemática é um operador que apresenta uma série de propriedades que são normalmente utilizadas para demonstrar relações na geoestatística. Seja c uma constante real e Z uma variável aleatória, temos que:

$$E(cZ) = cE(Z) \tag{2.4}$$

No caso do valor médio de uma variável aleatória ser constante, temos que a esperança matemática de um valor médio é o próprio valor médio. Essa propriedade é interessante ser ressaltada nos casos em que consideraremos a hipótese de estacionaridade de segunda ordem:

$$E(E(Z)) = E(c) = c \tag{2.5}$$

O operador esperança matemática é comutativo, isso significa que:

$$E(Y+Z) = E(Y) + E(Z) \tag{2.6}$$

Da mesma forma podemos definir o momento estatístico de segunda ordem centrado, também chamado de variância, como (2.7):

$$Var(Z) = E(Z - E(Z))^2 \tag{2.7}$$

Por uma simples transformação podemos definir a variância de uma variável aleatória como descrito na prova abaixo:

Demonstração. $Var(Z) = E(Z - E(Z))^2$
$Var(Z) = E\left(Z^2 - 2ZE(Z) + E(Z)^2\right)$
$Var(Z) = E(Z^2) - E(2ZE(Z)) + E(Z)^2$
$Var(Z) = E(Z^2) - 2E(Z)E(Z) + E(Z)^2$
$Var(Z) = E(Z^2) - 2E(Z)^2 + E(Z)^2$
$Var(Z) = E(Z^2) - E(Z)^2$ ■

O momento estatístico mais importante para a geoestatística é a covariância, sendo definida como o grau de dependência entre duas variáveis distintas Z e Y, por exemplo. Definimos a relação segundo a equação (2.8):

$$C(Z,Y) = E((Z - E(Z))(Y - E(Y))) \tag{2.8}$$

No caso em que as médias de Y e Z são iguais a covariância pode ser escrita como:

Demonstração. $E(Z) = E(Y) = m$
$C(Z,Y) = E((Z-m)(Y-m))$
$C(Z,Y) = E\left(ZY - Zm - Ym + m^2\right)$
$C(Z,Y) = E(ZY) - E(Zm) - E(Ym) + E(m^2)$
$C(Z,Y) = E(ZY) - mE(Z) - mE(Y) + E(m^2)$
$C(Z,Y) = E(ZY) - m^2 - m^2 + m^2$
$C(Z,Y) = E(ZY) - m^2$ ■

Essa demonstração é importante quando realizarmos as demonstrações da krigagem simples e ordinária posteriormente, utilizando a hipótese de estacionaridade de segunda ordem. A correlação e a variância são numericamente iguais quando consideradas as mesmas variáveis aleatórias. Se Y = Z podemos dizer que:

Demonstração. $C(Z,Y) = E((Z-E(Z))(Y-E(Y)))$
$C(Z,Z) = E((Z-E(Z))(Z-E(Z)))$
$C(Z,Z) = E(Z-E(Z))^2$
$C(Z,Z) = Var(Z)$ ■

Todas essas demostrações realizadas em uma variável aleatória qualquer também são aplicáveis nos casos das variáveis aleatórias regionalizadas.

2.6 A função variograma e a função covariograma

Como demonstrado na seção anterior podemos definir momentos estatísticos que representam características das variáveis aleatórias que definimos. Conceituamos a função covariância como sendo a esperança matemática do produto de duas variáveis aleatórias subtraído dos seus valores médios. Apesar de não estarmos interessados primordialmente na distribuição de probabilidades da nossa função aleatória é interessante definir o grau de correlação entre diversas variáveis aleatórias regionalizadas ao longo do nosso domínio. Definiremos agora a chamada função covariograma, que nada mais é que a correlação entre as variáveis aleatórias regionalizadas separadas de um vetor h.

A função covariograma é então definida como sendo a covariância entre a função aleatória separada de um vetor h entre as variáveis aleatórias regionalizadas consideradas. Logo temos que :

$$C(h) = E((Z(x) - m(x))(Z(x+h) - m(x+h))) \tag{2.9}$$

A função covariograma depende unica e exclusivamente do vetor considerado e não do suporte x das variáveis associadas. Em outros termos ela é dependente unicamente do fenômeno gerador do depósito e não da posição considerada para o seu cálculo.

Sob a hipótese de estacionaridade de segunda ordem podemos transformar a função covariograma na seguinte relação:

Demonstração. $m(x) = m \forall x \in D$
$C(h) = E((Z(x) - m)(Z(x+h) - m))$
$C(h) = E\left((Z(x)Z(x+h) - Z(x)m - Z(x+h)m + m^2)\right)$
$C(h) = E(Z(x)Z(x+h)) - m^2 - m^2 + m^2 = E(Z(x)Z(x+h)) - m^2$ ■

Outra função também importante a ser definida é o variograma. O variograma pode ser definido segundo a seguinte equação:

$$2\gamma(h) = E(Z(x+h) - Z(x))^2 \tag{2.10}$$

Sob a hipótese de estacionaridade de segunda ordem podemos transformar a função variograma na função covariograma tal como demonstrado na relação a seguir.

Denotaremos agora a variância à priori do fenômeno como sendo a covariância para um vetor de tamanho zero C(0)=Cov(Z(x),Z(x))=$\sigma^2(0)$.

Segundo a hipótese de estacionaridade de segunda ordem a variância de uma variável aleatória C(0)=c$\forall$ x $\in$ D:

Demonstração. $\gamma(h) = C(0) - C(h)$
$E\left(Z(x+h) - Z(x)\right)^2/2 = C(0) - E(Z(x)Z(x+h)) + m^2$

Devido a hipótese de estacionaridade de segunda ordem podemos decompor C(0) obtendo a seguinte relação:
$C(0) = E(Z(x)^2) - m^2$
$2C(0) = E(Z(x)^2) - m^2 + E(Z(x)^2) - m^2$
$2C(0) = E(Z(x+h)^2) - m^2 + E(Z(x)^2) - m^2$.: hipótese de estacionaridade

Isso nos leva à:
$E\left(Z(x+h) - Z(x)\right)^2 = E(Z(x+h)^2) - m^2 + E(Z(x)^2) - m^2 - 2E(Z(x)Z(x+h)) + 2m^2$
$E\left(Z(x+h) - Z(x)\right)^2 = E(Z(x+h)^2) + E(Z(x)^2) - 2E(Z(x)Z(x+h))$
$E\left(Z(x+h) - Z(x)\right)^2 = E(Z(x+h)^2 + Z(x)^2 - 2Z(x)Z(x+h))$
$E\left(Z(x+h) - Z(x)\right)^2 = E\left(Z(x+h) - Z(x)\right)^2$ ■

2.7 Valor de variograma médio

Como definido nas seções anteriores, o variograma é uma ferramenta estatística direcional, que depende exclusivamente do vetor associado e o fenômeno representado. Em alguns casos é interessante para nós definirmos um variograma médio, como sendo o valor da média dos variogramas dentro de um suporte determinado. Segundo a equação (2.11) definimos o variograma médio dentro de um suporte V contendo pontos nas posições x e y como abaixo:

$$\bar{\gamma}(V) = \frac{1}{VV'} \int_{V(x)} \int_{V(x')} \gamma(Z(x) - Z(x'))dxdx' \tag{2.11}$$

Em que x é um ponto no suporte V(x) e x' é um ponto no suporte V(x'). Na maioria das vezes a função integral não faz sentido no caso contínuo, já que calculamos o valor de $\bar{\gamma}$ a partir de pontos discretos em volume. Neste caso definimos sua forma discreta segundo a equação (2.12) para nxn' pontos situados dentro de um volume determinado.

$$\bar{\gamma}(V) = \frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} \gamma(Z(x_i) - Z(x_j)) \tag{2.12}$$

Estas equações serão importantes para definir questões de mudança de suporte nas estimativas. Em outras palavras, estimar em volumes e formas diferentes na geoestatística cria condições de variabilidade diversas. O valor médio de um painel de lavra subterrânea de algumas centenas de metros de extensão possui variabilidade totalmente diferente em relação à um bloco de um mina com algumas dezenas de metros apenas.

2.8 A variância de extensão

Consideremos o exemplo demonstrado na figura 2.2. Temos o ponto $Z(x_0)$ que desconhecemos e queremos determinar sua estimativa a partir dos pontos de 1 a 4 determinados ao seu arredor. Poderíamos apenas tirar a média destes valores, mas isso não levaria em consideração a correlação espacial entre as amostras e sua posição no espaço. Seria interessante também encontrar uma estimativa que possuísse o menor erro possível. A variância de extensão é a medida que demonstra quanto do erro tomamos por estimar um ponto a partir de uma combinação linear de amostras.

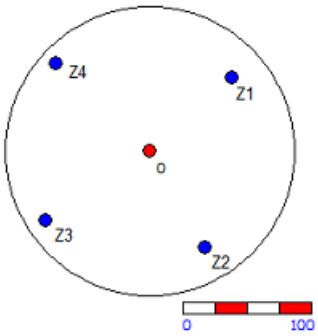


Figura 2.2: Ponto $Z(x_0)$ a ser estimado a partir de outros 4 pontos situados ao seu arredor

Considere o sinal * como sendo um valor estimado de uma variável aleatória. No ponto 0 o erro de estimativa pode ser demonstrado segundo a equação (2.13)

$$\varepsilon(Z^*(x_0)) = Z^*(x_0) - Z(x_0) \tag{2.13}$$

Podemos estimar o valor esperado de Z_0^* como sendo uma média ponderada dos valores amostrais de 1 a 4, logo temos que:

$$\varepsilon(Z^*(x_0) = \sum_{i=1}^{4} \lambda_i Z(x_i) - Z(x_0) \tag{2.14}$$

Para que o estimador não seja enviesado, devemos assumir que a esperança do erro de estimativa seja igual a zero. Considerando a hipótese de estacionaridade do valor médio, temos que:

Demonstração. $E(\varepsilon(Z^*(x_0))) = E(\sum_{i=1}^{4} \lambda_i Z(x_i) - Z(x_0)) = 0$
$\sum_{i=1}^{4} \lambda_i E(Z(x_i)) - E(Z(x_0)) = 0$
$\sum_{i=1}^{4} \lambda_i m - m = 0$
$\sum_{i=1}^{4} \lambda_i = 1$ ■

A variância do erro de estimativa pode ser então determinada segundo a prova abaixo :

Demonstração. $\sigma^2(\varepsilon(Z^*(x_0))) = E\left(\sum_{i=1}^{4} \lambda_i Z(x_i) - Z(x_0)\right)^2$
$= E\left(\sum_{i=1}^{4}\sum_{j=1}^{4} \lambda_i \lambda_j Z(x_i)Z(x_j) - \sum_{j=1}^{4} 2\lambda_i Z(x_i)Z(x_0) + Z(x_0)^2\right)$
$= \sum_{i=1}^{4}\sum_{j=1}^{4} \lambda_i \lambda_j E(Z(x_i)Z(x_j)) - \sum_{i=1}^{4} 2\lambda_i E(Z(x_i)Z(x_0)) + E(Z(x_0)Z(x_0))$

Adicionando e retirando o valor da média m no problema temos:

$= \sum_{i=1}^{4}\sum_{j=1}^{4}\lambda_i\lambda_j E(Z(x_i)Z(x_j)) - m^2 - \sum_{i=1}^{4} 2\lambda_i E(Z(x_i)Z(x_0)) + 2m^2 + E(Z(x_0)Z(x_0)) - m^2$
$= \sum_{i=1}^{4}\sum_{j=1}^{4}(\lambda_i\lambda_j E(Z(x_i)Z(x_j)) - m^2) - \sum_{i=1}^{4} 2(\lambda_i E(Z(x_i)Z(x_0)) - m^2) + (E(Z(x_0)Z(x_0)) - m^2)$
$= \sum_{i=1}^{4}\sum_{j=1}^{4}\lambda_i\lambda_j Cov(Z(x_i),Z(x_j)) - \sum_{i=1}^{4} 2\lambda_i Cov(Z(x_i),Z(x_0)) + Cov(Z(x_0),Z(x_0))$

■

A variância do erro de estimativa, também chamada de variância de extensão, denotada como σ_e^2 é composta de três partes, uma com a autocovariância entre as amostras a serem utilizadas, a covariância entre o ponto estimado e as amostras e a variância do ponto estimado.

2.9 Krigagem

O termo krigagem foi criado em homenagem a Daniel G. Krige que iniciou os trabalhos de geoestatística nas minas de ouro na África do Sul. Basicamente a krigagem nada mais é do que encontrar os valores dos ponderadores que minimizam a variância de estimativa, logo podemos definir a krigagem como sendo o conjunto de equações determinado por (2.15):

$$\frac{\partial}{\partial\lambda_i}\sigma_e^2 = 0|\forall i \tag{2.15}$$

Utilizando a demonstração da seção anterior conseguimos verificar que o sistema de krigagem leva às seguintes equações. Para isso basta expandir a série para cada amostra considerada e derivar parcialmente cada um dos valores (2.16):

$$\sum_{i=1}^{n}\lambda_i Cov(Z(x_i),Z(x_j)) = Cov(Z(x_i),Z(x_0))|\forall j \tag{2.16}$$

Diversos tipos de krigagem são determinadas a partir desta solução simples. Cada um dos sistemas possuí restrições quanto as hipóteses de cada modelo tornando o sistema de equações parciais em um problema de minimização com restrições, utilizando o operador lagrangiano, tal como demonstrado em (2.17), sabemos que o gradiente do erro de estimativa deve ser ortogonal à restrição indicada.

$$\nabla\sigma_e^2 \perp \mu R' \tag{2.17}$$

Tal que R é uma restrição dada aos pesos no sistema de krigagem e μ é o operador lagrangiano. R' é portanto a derivada parcial da restrição em relação aos ponderadores. Essa restrição pode estar associada à uma condição de não viés do estimador ou de outra hipótese estabelecida.

2.10 A variância de dispersão

Uma das maiores contribuições da geoestatística para a mineração e para efeitos de engenharia foi o conhecimento sobre a variabilidade em escala dos fenômenos espaciais. Na maioria das vezes estamos interessados em determinar o valor esperado de uma variável aleatória, sendo que esta pode apresentar uma dispersão, ou seja, o valor médio pode se situar dentro de limites desconhecidos.

É bem intuitivo pensarmos que o valor médio estimado tende a depender do suporte considerado. Um volume muito grande de um painel de uma mina subterrânea estimado com seis ou sete furos de sondagem é bem mais preciso que o valor estimado de um bloco de minério de 10cm de comprimento.

A variância de dispersão é uma medida de quanto um volume a ser estimado pode variar segundo uma informação de um suporte diferente pode fornecer. Observe a figura 2.3. No item A temos o valor médio das amostras contidas dentro do bloco como sendo uma estimativa daquele volume. A variância dos dados é tomada como sendo a variância das amostras A pelo valor médio dentro daquele suporte. A média de diversos blocos diferentes, como demonstrado em B relata a variância de dispersão. Neste caso temos uma medida qualitativa de quanto a informação de um suporte pode influenciar na estimativa de outro.

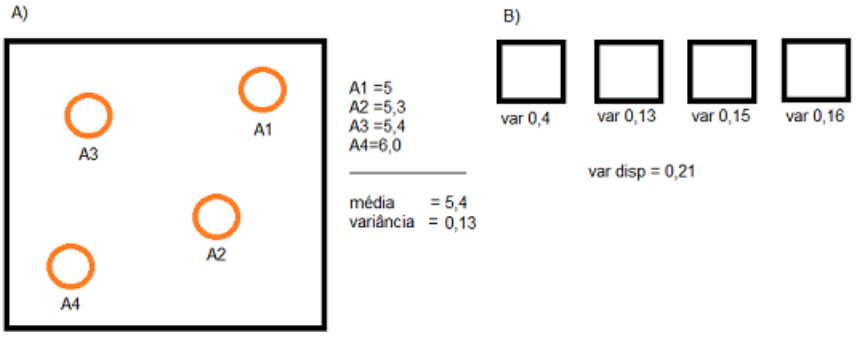


Figura 2.3: Exemplo para variância de dispersão A) Um bloco sendo estimado com valores de amostra laranja contido dentro dele b) Variância de dispersão como a média dos valores de variância para cada bloco

Definimos a variância de dispersão de um suporte v dentro de um suporte V como definido na equação (2.18)

$$D^2(V/v) = E[S^2(Z_v(x))] = E\left[\frac{1}{n}\sum_{i=1}^{n}(Z_v(x_i) - Z_V(x))^2\right] \tag{2.18}$$

Note que Z_V não necessariamente precisa ser a média de valores contidos dentro do suporte V.

Podemos utilizar qualquer estimativa para determinar o valor da variância de dispersão, podendo ser utilizada até mesmo o valor krigado, em que veremos em seções posteriores. A variância de dispersão é uma medida da entropia da informação e o quão ela influencia na estimativa em um dado suporte.

2.11 Exercícios

Exercise 2.1 Enumere em uma lista todas as variáveis aleatórias regionalizadas que você possui em seu objeto de estudo. Indique ao lado se elas são somáticas ou não. Ex.: Teor-> somático, Condutibilidade hidráulica -> não somático. ■

Exercise 2.2 Cinco ações de uma mineradora possuem rentabilidade de 5, 10,20,4 e 5 Unidades monetárias. Se a probabilidade de renda destas ações forem iguais a 40%, 35%, 10%, 10% e 5% qual é o valor esperado para a renda de todas as ações. Resp.:8.15 UM ■

Exercise 2.3 Cinco amostras possuem valor de teor iguais a 2%, 2.5%,2.3%,2.1% e 2.7%. Se o volume das amostras é de 5,4,3,5 e 7 cm^3 qual é o teor médio das amostras. Resp.: 2,34% ■

Exercise 2.4 Prove que o valor do resíduo da função aleatória é ortogonal à sua tendência, ou seja $Cov(R,m) = 0 \forall x \in D$ sendo D o domínio do depósito. ■

Exercise 2.5 Prove que a covariância de duas variáveis aleatórias independentes seja igual a zero. Dica.: Tome o valor de $E(XY) = E(X)E(Y)$

■

3. Estatística univariada

O primeiro passo para a avaliação de um recurso mineral é descrever as amostras retiradas em campo. O que desejamos nesta etapa é resumir estatisticamente um conjunto grande de informações. Uma tabela contendo vários números é de difícil compreensão quando lida, mas ao resumir a informação relatando que 90 por cento dos dados estão acima de um teor de 5g/tonelada, ou que 70 por cento do depósito mineral é constituído do litotipo 1, torna-se mais fácil a tarefa da tomada de decisão. Dados são convertidos em informações quando é possível tomar juízos a partir deles.

Nete caso estamos interessados em saber se as amostras estão acima de um valor de cut-off prescrito, quais são os valores mínimos e máximos, se suas variações são pequenas ou grandes, quais são os valores mais comuns e incomuns. Todas essas informações são necessárias para que um engenheiro ou geólogo possa resolver problemas em uma mina.

Imaginem que um avaliador tenha em mãos os dados de testemunhos de sondagem regularmente espaçados ao longo de uma área, com um valor de teor médio de 5.2% e um desvio padrão de 2%. Ou seja, há uma grande chance de que os valores de teor da área estejam entre 7.2% e 3.2%. Para aquele depósito mineral, o avaliador sabe que o cut-off para o depósito é de 12%. Dessa forma as chances daquela área se tornar um recurso são ínfimas. Ele pode recusar a campanha de exploração se não houverem recursos para continuá-la. Os dados se transformaram em uma informação a partir das estatísticas e com ela pode-se tomar uma decisão sobre o empreendimento.

É importante entender que toda a estatística transforma dados em informação, mas em contra-partida perde a sensibilidade dos valores das amostras. A média e o desvio padrão são funções que resumem os dados, mas deixam de ressaltar peculiaridades inerentes da distribuição dos dados.

Tomemos como exemplo dois blocos dentro de uma mina. Digamos que um seja um minério com alto valor agregado, e outro um estéril muito pobre. Digamos que a média dos seus valores ainda seja considerada minério, tal como mostrado na figura 3.1. Se lavrarmos dois blocos conjuntamente temos ainda um grande bloco de minério, pois sabemos que seu valor médio está acima do limite econômico, no entanto, perdemos a sensibilidade dos dados pois sabemos que há um valor que poderia ser recusado como um estéril.

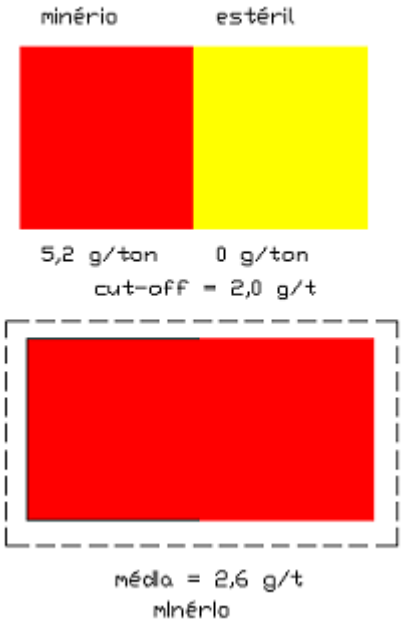


Figura 3.1: Média de dois blocos - Um estéril e outro minério. Os dois blocos são considerados conjuntamente como minério se tomarmos o valor médio

Esta falta de sensibilidade dos dados, que as estatísticas causam, faz com que seja necessário incorporar mais funções na análise para entendermos as propriedades dos dados. É inadmissível caracterizarmos um depósito apenas situando seus valores médios. É sempre importante agregarmos o máximo de informação possível.

3.1 Valores outliers

A primeira etapa da geoestatística é a validação das amostras. Devemos antes de tudo verificá-las para que não encontremos valores discrepantes (outliers) ou incoerências nos dados. Além disso, a estatística descritiva é responsável por determinar as primeiras informações a respeito do depósito mineral. Esta é uma etapa necessária antes que se prossigam com os métodos geoestatísticos. É obrigação do avaliador de reservas verificar se as metodologias de amostragem e os valores inseridos no banco de dados estão corretos. A Tabela 3.1 é um exemplo de como valores anômalos podem aparecer. Nota-se claramente que as amostras 1 e 3 estão erradas. Primeiramente porque não existem valores de teor percentuais acima de 100% e também porque não existem teores descritos como letras. No entanto, a amostra 4 também está errada, porque o minério composto por limonita não pode apresentar um valor de teor de ferro de 72%, pois é incompatível com a química da mineralogia.

Tabela 3.1: Tabela de teores do minerio de ferro

Índice	Minério	Teor(%)
1	Hematita compacta	120%
2	Hematita granular	53%
3	Magnetito	0.i3
4	Limonita	72%

Um gráfico que auxilia para determinar valores outliers é chamado de boxplot. Ele demonstra a disposição dos dados em um eixo e limita os valores das amostras em uma caixa contendo os quartis das amostras. Os valores que se situam acima ou abaixo de 1.5 do intervalo interquartil representam outliers. O intervalo interquartil é também determinado como a diferença entre os valores do terceiro quartil e do primeiro quartil. A figura 3.2 demonstra o gráfico boxplot e suas dimensões.

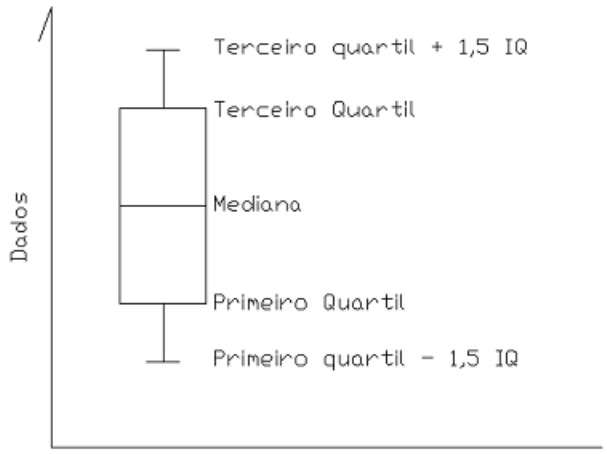


Figura 3.2: Representação de um gráfico de caixa dividida entre os intervalos das amostras

Os valores anômalos ou outliers são demonstrados na figura 3.3 como pontos circulados fora das barras que representam os limites de aceitação dos valores da amostra.

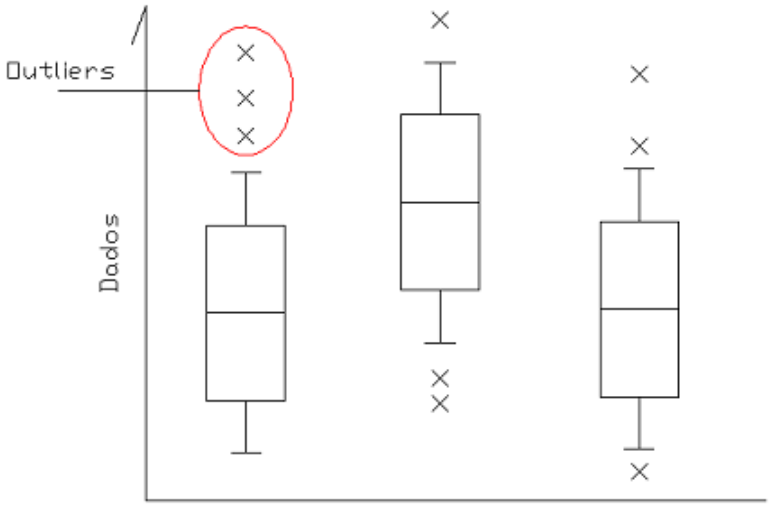


Figura 3.3: Representação dos valores outliers no gráfico boxplot - Pontos circulados em vermelho

É importante entender que os dados anômalos nem sempre são valores errados. Eles podem ser valores reais representantes de uma anomalia da natureza. Poderíamos encontrar, por exemplo, em um depósito de ouro uma pepita com um valor agregado muito alto, mas apesar de ser um dado correto ele não representa o conjunto de amostras como um todo.

O tratamento de dados discrepantes ou outliers pode ser realizado de diversas maneiras.

- Podemos retirar os dados da avaliação. Esta exclusão deve ser criteriosa e se basear na

qualidade da amostra, na recuperação do testemunho de sondagem ou em um critério definido. A amostragem é a etapa mais cara no processo de avaliação de reservas e por isso a retirada de valores outliers não pode ser realizada como uma rotina.

- Podemos ponderar o valor da amostra durante a análise. Se um valor outlier representar um dado discrepante em meio a um conjunto de amostras com baixo valor, podemos reduzir sua influência ponderando as estatísticas pela sua área de influência.
- Podemos tratar a amostra como uma classe separada das outras amostras. Diferenciando as amostras outliers do resto podemos indicar regiões que apresentam propriedades estatísticas diferentes.

É importante salientar que qualquer método de estimativa não cria informação. Se os dados descritivos de uma análise inicial do depósito mineral não indicarem recursos adequados a estimativa constatará da mesma forma a informação.

Qualquer método de inferência não extrapola os valores mínimos e máximos de um depósito mineral. Descrever é antes de tudo um passo que necessita encontrar propriedades de algo. A descrição deve conter os aspectos mais importantes de um depósito mineral, tal como mínimo e máximo encontrados, valores médios, dispersão. Da mesma forma que desenhar é uma atividade altamente explicativa para descrever um problema, as estatísticas gráficas desempenham papel fundamental na avaliação inicial.

3.2 Descrição espacial das amostras

O primeiro passo para a descrição espacial é verificar a disposição geométrica das amostras. A figura 3.4 demonstra um depósito polimetálico de Jura. O atributo é o tipo de rocha de um dado período geológico.

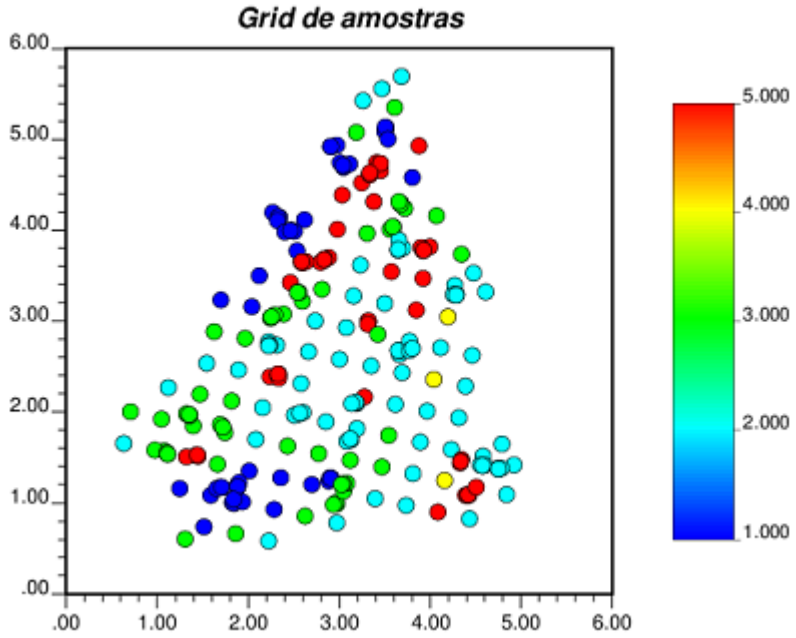


Figura 3.4: Disposição das amostras no espaço. Cores diferenciadas mostrando tipos de rocha em períodos geológicos diferentes

Podemos ver que as amostras estão dispostas de forma irregular em um formato de delta de um rio. A orientação do tipo de rocha 1 se encontra ao oeste e parte ao sul, enquanto a do tipo 5 se encontra distribuído mais ao norte. Qualquer estimativa realizada a partir desta configuração de

amostras deve respeitar os valores iniciais. Se por exemplo, iniciássemos uma explotação cujo o interesse seria o litotipo 1, provalvelmente começaríamos a retirar o material de oeste para leste para reduzir o fluxo de caixa do empreendimento.

A 3.5 demonstra o atributo Cádmio acerca deste depósito. Note que o depósito é bem homogêneo quanto a distribuição do elemento.

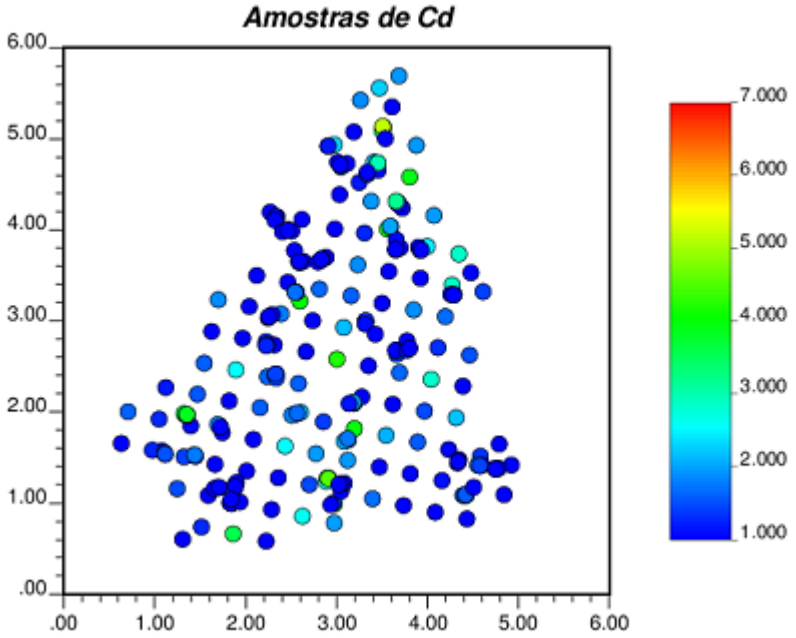


Figura 3.5: Disposição do Cd

As informações espaciais das amostras garantem alternativas para associar as variáveis com suas componentes genéticas e interpretar seus valores segundo objetivos desejados. Notamos que o litotipo 1 parece ter maior correlação com valores baixos do teor de Cádmio do que o litotipo 2, que parece ter correlação com valores um pouco mais altos. O primeiro passo da análise visual é verificar regiões de risco, de valores mais ricos ou pobres e identificar padrões de continuidade nas amostras.

3.3 Histograma

A descrição das estatísticas das amostras é uma forma inicial para aglomerar um conjunto de informações extensos. Um gráfico de grande utilidade para verificar a frequência dos dados é o histograma. Este é uma figura de barras em que a altura de cada retângulo representa a frequência de uma classe.

A 3.6 representa um histograma da variável Cádmio do depósito de Jura.

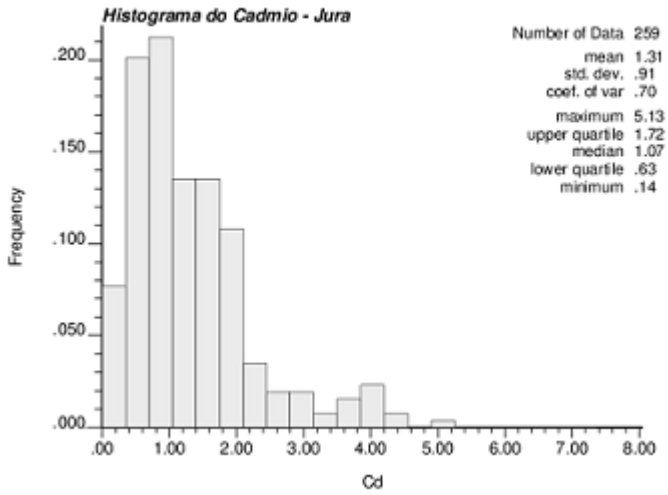


Figura 3.6: Histograma do Cd

Na 3.7 podemos ver que a classe de teores de $0,04$ a $0,75 g/ton$ ocupa uma proporção de 20 % dos dados.

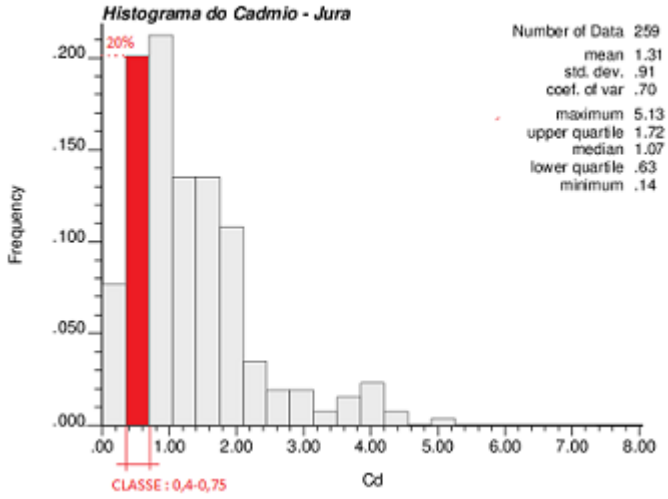


Figura 3.7: Histograma do Cd - Classe marcada

Ao contrário da análise espacial, no histograma estamos apenas interessados em determinar o conjunto de informações das amostras. Perdemos a noção de suporte quando utilizamos esta estatística, porém ganhamos recurso para o entendimento valores das amostras.

Outra forma de representar um histograma é na sua forma acumulada. A figura 3.8 é uma demonstração do gráfico acumulado.

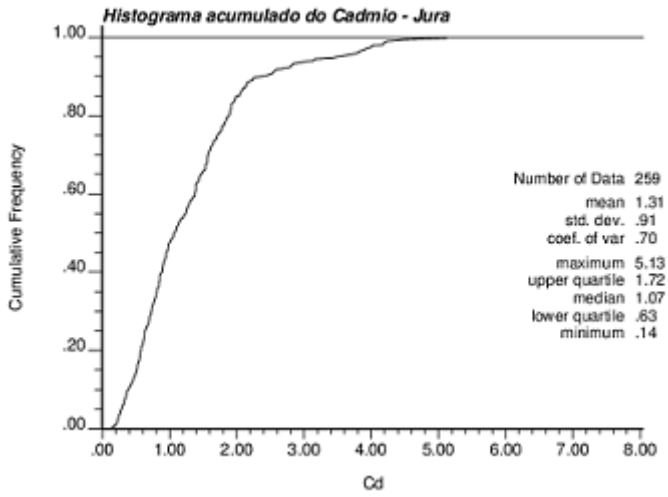


Figura 3.8: Histograma do Cd acumulado

A figura 3.9 demonstra a leitura do gráfico acumulado.Podemos notar por este gráfico que 60 por cento dos valores estão abaixo do teor de $1,5 g/tonelada$.

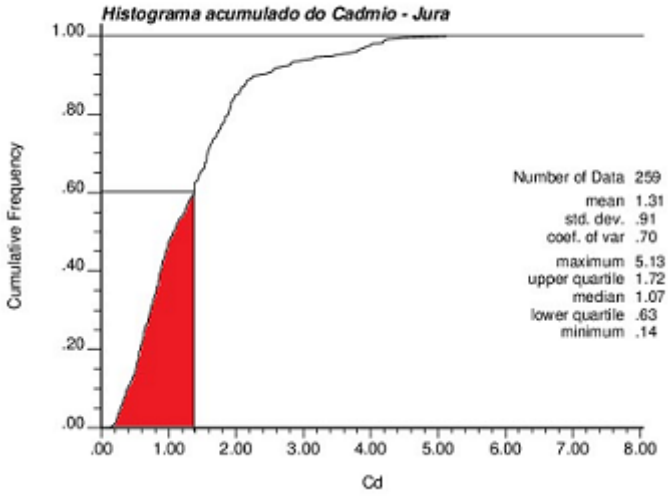


Figura 3.9: Histograma do Cd acumulado - Leitura

O formato do histograma também é um importante parâmetro para a inferência de distribuições de probabilidade. A partir dele podemos visualizar uma possível distribuição de probabilidade e dar um "chute"para testarmos se esta se encaixa na distribuição das amostras.

É importante para a escolha da resolução do gráfico um número de classes condizentes. Muitas classes farão várias barras terem valores de apenas 1 unidade, enquanto apenas uma classe incorporará todas as amostras. A fórmula de Sturges é uma boa estimativa para o número mínimo de classes para o histograma, mas nada compete com a iniciativa do analista em encontrar a quantidade que mais se adéqua ao problema. É importante que o histograma demonstre claramente o comportamento de simetria da distribuição, dos valores mais frequentes e da dispersão da variável.

A simetria é uma medida que indica quão próximos estão os valores distribuídos de um conjunto

de amostras perto do seu valor médio. O histograma da figura 3.6 demonstra uma distribuição claramente assimétrica em que os valores mais baixos possuem proporções maiores que os valores mais altos. Isso significa que neste depósito a chance de encontrar uma amostra com teor de Cádmio baixa é alta, mas no entanto, ainda é possível encontrar um valor muito mais alto em uma frequência menor. Distribuições simétricas geralmente são mais comportadas. Estimar seus valores e determinar seus parâmetros se torna mais simples, porque elas giram entorno de um valor esperado. Espera-se menos "surpresas"quando uma distribuição é simétrica.

3.4 Estatísticas pontuais

Outra forma de resumir e descrever os dados é através de estatísticas pontuais. Elas resumem a informação do conjunto de amostras em uma única medida descrevendo-o como um todo. Se fôssemos comparar a descrição pontual com o retrato falado de um criminoso, cada estatística seria apenas uma parte do rosto, a média o nariz e a variância as orelhas, por exemplo.

Parece um pouco tolo este tipo de descrição, no entanto, as características de um rosto humano são de fácil reconhecimento para qualquer indivíduo. Isso motivou a criação de um tipo de gráfico específico na estatística chamado também de rostos de Chernoff.

Existem diversas estatísticas pontuais, cada uma medindo uma propriedade das amostras, mas apenas estamos interessados em 4. Medidas de tendência central, medidas de dispersão, medidas de assimetria e de achatamento.

É importante salientar que apenas uma estatística pontual não é uma medida que garante informação completa a respeito de um conjunto de dados.

Um depósito mineral pode ter valor médio de 50g de ouro por tonelada, enquanto outro tenha 45g de ouro por tonelada, e ainda assim o segundo depósito seja mais rico.

Isso acontece porque as medidas pontuais de tendência central como a média devem estar sempre associadas com uma medida de dispersão. Se o depósito de 50 g por tonelada possuir uma menor dispersão, e o depósito de 45 g/ton possuir uma maior, para um dado cut-off o depósito de 45g/ton pode ser mais rico.

A Figura (3.10) demonstra esta situação graficamente. Notamos que a a distribuição A, apesar de possuir uma média menor que a distribuição B, ainda assim relata um depósito mais rico para o cut-off considerado.

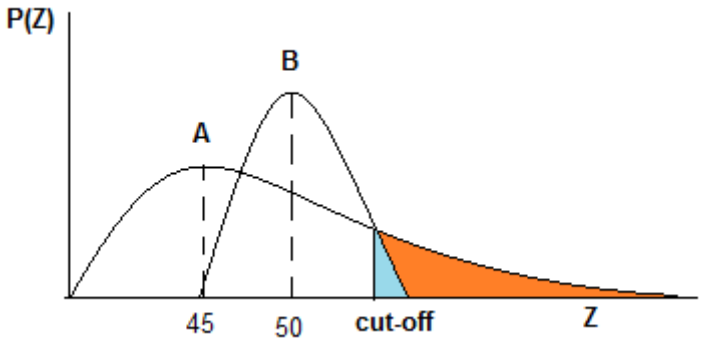


Figura 3.10: Exemplo de duas distribuições A e B relatando um depósito mais rico A com média menor que B. Área azul mostrando a contribuição da distribuição B acima do cut-off e área laranja mostrando a contribuição de A acima do cut-off

3.4.1 Medidas de tendência central

As medidas de tendência central são estatísticas calculadas a partir das amostras que representam o centro de massa do conjunto. Analogamente ao ponto de equilíbrio de uma barra, estas representam o centro de dispersão dos dados.

Note que esta é uma convenção matemática. O valor médio não representa necessariamente um valor do conjunto de amostras e nem tão pouco pode representar um valor mais provável, mas apenas um centro da dispersão dos dados.

Como exemplo, a média do lançamento de um dado é 3.5 que não constitui nem um valor possível dos dados, nem ao menos o mais provável. Distribuições com histogramas mais simétricos e comportados geralmente podem associar o valor mais provável com a média.

As medidas de tendência central mais comuns são a média aritimética, a moda, a média ponderada e a mediana. A média aritimética pode ser descrita segundo a equação (3.1) em que x são os valores das amostras e n o número de amostras.

$$\bar{x} = \sum_{i=0}^{n} x_i \tag{3.1}$$

A moda pode ser descrita como o valor mais frequente observado no conjunto de amostras. Nem sempre uma distribuição pode apresentar um valor de moda. Para variáveis contínuas, muitas vezes temos apenas um valor de cada realização. Neste caso não temos moda nenhuma ou todos os valores são a moda e não faz sentido definí-la.

A média ponderada pode ser descrita pela equação (3.2)

$$\bar{x} = \sum_{i=0}^{n} \lambda_i x_i \tag{3.2}$$

Em que λ_i representa o ponderador de cada realização. O ponderador é o valor de cada peso dividido pela soma de pesos totais (3.3)

$$\lambda_i = \frac{p_i}{\sum_{i=1}^{n} p_i} \tag{3.3}$$

Logo sabemos que a soma de todos os ponderadores de uma média ponderada é igual a 1 sempre. A média ponderada dos teores é um exemplo em que temos o valor do peso igual ao volume da amostra e o ponderador como a relação de cada volume pelo volume total das amostras.

3.4.2 medidas de posição

Os quartis são medidas de posição que indicam o valor percentual de uma série ordenada de dados. O primeiro quartil indica o valor ao qual os dados se dividem em 25 por cento de ordem crescente, o segundo quartil também chamado de mediana, divide os dados em 50 por cento e o terceiro quartil em 75 por cento. Outros valores de posição são os percentis, que representam o valor para um percentual acumulado das amostras.

Se obtivermos um conjunto de dados iguais a $50, 34, 27, 54, 25, 43, 15, 12$ contendo 8 valores então podemos ordená-los em crescente de tal forma que teremos $12, 15, 25, 27, 34, 43, 50, 54$. O valor do primeiro quartil será, segundo os dados ordenados, 15. O terceiro quartil será 43. E a mediana será igual a 27.

3.4.3 medidas de dispersão

Outras medidas importantes são as de dispersão. Entre as mais comuns podemos citar a variância, o desvio padrão e os valores de mínimo e máximo. A variância pode ser descrita pela equação (3.4)

$$s^2 = \frac{\sum_{i=0}^{n} (x_i - \bar{x})^2}{n-1} \tag{3.4}$$

Em que n-1 é o número de graus de liberdade da amostra, tal que este pode ser definido pelo número de amostras menos o número de estatísticas utilizadas durante o cálculo. Note que para a operação da variância precisamos antes determinar o valor da média. É uma medida que não apresenta as mesmas unidades que a das amostras, para isso geralmente utilizamos o desvio padrão, que pode ser calculado como a raiz quadrada dos valores da variância.

3.4.4 Conjugando estatísticas pontuais

Como dito anteriormente, é sempre importante conjugar estatísticas pontuais diferentes de forma a garantir a melhor informação possível. Uma destas alternativas é adicionar ao valor médio um número de desvios padrões de forma a garantir que um conjunto de dados esteja situado dentro destes limites. Para isso utilizaremos uma das mais renomadas relações estatísticas.

A desigualdade de Chebyshev é uma identidade que implica em um valor mínimo de probabilidade para que uma realização esteja dentro de um intervalo múltiplo do desvio padrão. Podemos definir a equação (3.5) como a desigualdade de Chebyshev.

$$P(|Z - \mu| \geq k\sigma) \leq 1/k^2 \tag{3.5}$$

Em que Z é o valor da variável aleatória, μ é o valor da média da população, σ é o valor do desvio padrão da população e k é uma constante proporcional. A desigualdade de Chebyshev é independente do valor da distribuição de probabilidades.

Para um k igual a 2, sabemos que existe uma probabilidade de no mínimo 75 por cento de que o valor da amostra esteja em dois desvios padrões da média.

Podemos caracterizar as amostras então por uma medida de posição e de dispersão conjuntamente. Ao descrever as amostras é bem claro que devemos associar no mínimo dois de seus parâmetros, como por exemplo, dizer que as amostras de teor de ouro possuem valores entre $(50 \pm 20)\, ppm$ em que 20 representaria dois desvios padrões de 10 ppm e 50 ppm seu valor médio.

3.4.5 Assimetria

Outra medida pontual importante também é a assimetria. Esta se caracteriza pela diferença de proporções de uma distribuição de amostras segundo ao redor de seu valor mais frequente.

A figura (3.11) demonstra a distribuição de dados assimétrica. O item a) representa uma distribuição assimétrica positiva, enquanto o item b) representa uma distribuição assimétrica negativa. A assimetria positiva é caracterizada por um valor da mediana abaixo do valor médio, enquanto a assimetria negativa se caracteriza por uma alta proporção de valores altos.

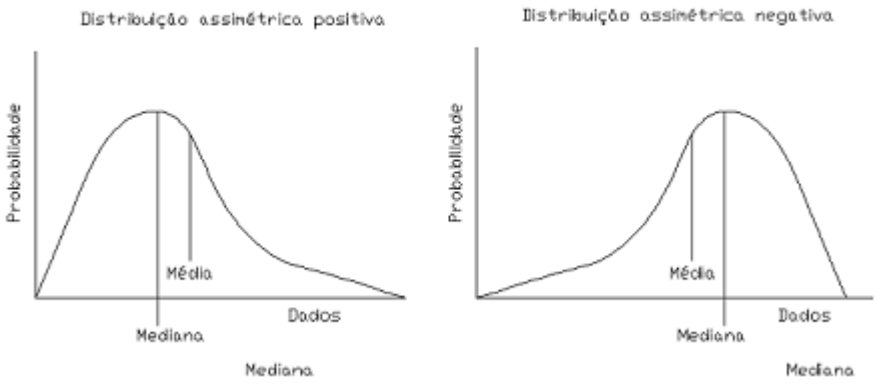


Figura 3.11: Assimetria de uma distribuição de dados a) Assimetria positiva b) assimetria negativa

Uma das medidas de assimetria mais comuns é o coeficiente de Pearson que pode ser expresso pela equação (3.6)

$$\varepsilon = (\bar{x} - m_0)/s \tag{3.6}$$

Em que m_0 é a moda dos dados, $\bar{x}$ é o valor médio das amostras e s é o desvio padrão das amostras.

Distribuições com característica de assimetria positiva são muito comuns na avaliação de depósitos minerais, principalmente no tratamento de commodites erráticos tal como ouro e diamante. Nesses depósitos podem ocorrer anomalias raras e uma amostra constituir em alto valor. Esta propriedade também é chamada de efeito pepita e será melhor tratada no capítulo de Continuidade espacial.

3.4.6 Coeficiente de variação

Em certos momentos é importante comparar variáveis aleatórias de tipos diferentes. Para sabermos se uma distribuição é mais errática que outra, neste caso, não bastaríamos comparar seus valores de variância. Valores que possuam médias maiores tendem a apresentar dispersões também maiores. Para isso utilizamos o coeficiente de variação, que nada mais é do que o desvio padrão de uma distribuição pelo seu valor médio. Desta forma "igualamos"diferentes distribuições em um único coeficiente comparativo.

O coeficiente de variação pode ser dado pela equação (3.7)

$$CV = \frac{s}{\bar{x}} \tag{3.7}$$

3.5 Inferência Estatística

Após analisados os dados amostrais podemos utilizar funções para modelar populações dos dados. É importante notar que neste caso estamos definindo uma lei de probabilidade para todo o depósito, independente dos valores locais e ainda não estamos utilizando a geoestatística propriamente. É de pouco interesse para nós modelarmos um depósito mineral sem o conhecimento do suporte dos dados. Logo esta etapa demonstrada aqui, como uma estatística univariada, é muito mais importante a ser realizada depois dos métodos de estimativa ou durante os procedimentos de transformação dos dados. Vamos iniciar alguns tipos de distribuições mais importantes. ,

3.5.1 Famílias de distribuições estatísticas

Uma função de densidade de probabilidade de uma variável aleatória nada mais é do que uma função $p(X = x)$ que correlaciona cada realização da variável aleatória X a uma dada probabilidade. Como consequência da definição algumas condições estão associadas:

- $p(x) \leq 1 \forall x$
- $\int_{-\infty}^{\infty} p(x)dx = 1$ para distribuições contínuas
- $\sum_{x=-\infty}^{\infty} p(x) = 1$ em que a e b são limites para a distribuição discreta

Distribuição de Poisson

Esta é uma distribuição discreta amplamente utilizada para experimentos ditos de eventos "raros", ou seja, utilizada para modelar eventos que a probabilidade de ocorrência é diretamente proporcional ao tempo de espera.

Em filas de caminhões, por exemplo, é muito comum a utilização da função de distribuição de Poisson para medir a probabilidade de chegada de um equipamento, pois é de se esperar que para um pequeno intervalo de tempo após a saída de um caminhão da frente de lavra, a probabilidade da chegada de outro seja pequeno. Outro exemplo é a frequência de fraturas em uma rocha. É de se esperar que para tamanhos pequenos a quantidade de fraturas seja pequena, enquanto para tamanhos grandes de rocha essa densidade aumente.

Evidentemente todo o processo de modelagem das distribuições utilizadas nas simulações deve advir antes de tudo de uma amostragem, sendo que suposições podem causar incoerências no modelo. A função de distribuição de Poisson pode ser escrita segundo a equação (3.8)

$$P(X = x) = \frac{\exp^{-\lambda} \delta^x}{x!} \tag{3.8}$$

Em que x é o evento da variável, P(X =x) é a probabilidade associada àquele evento e $\lambda = E(X)$ sendo o parâmetro da função. Tal como qualquer distribuição de probabilidades sabemos que a soma de todos os eventos possíveis deve gerar um resultado igual a 1. Podemos demonstrar isso de acordo com a prova

Demonstração. Sabendo que a função exponencial pode ser aproximada por uma série de Taylor como a seguir temos :

$e^{\lambda} = \sum_{n=0}^{\infty} \frac{\lambda^x}{x!}$
Então:

$\sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{\exp^{-\lambda} \lambda^x}{x!}$
$\sum_{x=0}^{\infty} P(X = x) = \exp^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$
$\sum_{x=0}^{\infty} P(X = x) = \exp^{-\lambda} \exp^{\lambda} = 1$ ■

Distribuição Gaussiana

Esta talvez seja uma das funções de densidade de probabilidade mais populares e representa um grande papel na estatística. A distribuição é um modelo simétrico e descrito por dois parâmetros, a média da população e a variância. A função de densidade de probabilidade da distribuição pode ser desrita segundo a equação (3.9)

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma}} exp^{\frac{(-x-\mu)^2}{2\sigma^2}} \tag{3.9}$$

Em que σ^2 é a variância da distribuição aleatória e μ é a média. O caso particular da distribuição gaussiana é quando sua média é igual a zero e variância é igual a 1, neste caso temos uma distribuição padronizada segundo a equação (3.10)

$$P(X = x) = \frac{1}{\sqrt{2\pi}} exp^{\frac{(-x)^2}{2}} \tag{3.10}$$

Uma variável aleatória pode ser padronizada segundo a relação (3.11)

$$X_p = (X - \mu)/\sigma \tag{3.11}$$

Que nada mais é do que uma operação de deslocamento da variável aleatória pela sua média e encurtamento da distribuição pelo seu desvio padrão.

Para demonstrar que a distribuição gaussiana possui soma de todos os seus eventos igual a 1 devemos antes lembrar que ela é uma distribuição simétrica, logo a soma dos valores à esquerda do valor médio da distribuição é idêntico à soma dos valores à direita da distribuição. Logo temos a relação (3.12) :

$$\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} exp^{\frac{(-x)^2}{2}} dx\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} exp^{\frac{(-t)^2}{2}} dt \int_{-\infty}^{\infty} exp^{\frac{(-u)^2}{2}} du \tag{3.12}$$

Que é o mesmo de:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} exp^{\frac{(-x)^2}{2}} dx = \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \int_{0}^{\infty} exp^{\frac{-(t^2+u^2)}{2}} dtdu \tag{3.13}$$

Logo temos a seguinte prova abaixo:

Demonstração. $\frac{1}{\pi} \int_0^\infty \int_0^\infty exp^{\frac{-(t^2+u^2)}{2}} dtdu =$
Essa relação pode ser transformada em coordenadas polares tal que:
$\frac{1}{\pi} \int_0^\infty \int_0^{2\pi} rexp^{\frac{-r^2}{2}} drd\theta =$
$2\pi \frac{1}{\pi} \int_0^\infty rexp^{\frac{-r^2}{2}} dr$
$2\lim_{y\to\infty} -\frac{e^{\frac{-r^2}{2}}}{2} \Big|_0^y = 1$

■

Distribuição Lognormal

A distribuição lognormal é uma distribuição assimétrica e positiva, geralmente associada na mineração com depósitos de elementos raros , tais como ouro, diamante e platina. Pode ser considerada uma distribuição cujo logaritmo é normalmente distribuído. A equação (3.14) demonstra a função de densidade de probabilidade para a distribuição lognormal.

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma}} \frac{1}{x} exp^{\frac{(-log(x))^2}{2}} \tag{3.14}$$

O Valor esperado da distribuição pode ser demonstrado segundo a equação (3.15)

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \tag{3.15}$$

Estimando a média da população

O processo de inferência estatística resume-se em determinar características da população a partir de dados amostrais. Depois de determinado o histograma da variável aleatória considerada podemos observar o seu padrão de distribuição (assimétrico/simétrico) e seus valores de frequência e determinar uma possível distribuição para os valores da população. Um bom estimador para a média da população é, por exemplo a média das amostras. Considere um conjunto de n amostras Z, logo temos segundo a equação (3.16)

$$E(Z_(x)) = E\left(\sum_{i=1}^{n} Z(x_i)\right)/n = \left(\sum_{i=1}^{n} E(Z(x_i))\right)/n = \left(\sum_{i=1}^{n} m\right)/n = m \tag{3.16}$$

Ou seja, sob a hipótese de estacionaridade de segunda ordem podemos considerar que a média das amostras é um bom estimador para a média da população ou do depósito mineral.

Enquanto a variância no entanto temos segundo a equação (3.17)

$$Var(\mu) = Var\left(\sum_{i=1}^{n} Z(x_i)/n\right) = \sum_{i=1}^{n} 1/n^2 Var\left(Z(x_i)\right) = \sigma^2/n^2 \tag{3.17}$$

Em outros termos, sob a hipótese de estacionaridade, a variância da média populacional tende a reduzir de acordo com o número de amostras tomadas. Isso também é chamado de efeito de suporte, pois quanto mais informações temos com a amostragem, mais o valor esperado de uma função aleatória tende a ser o correto. A figura (3.12) demonstra como o valor médio tende a cada vez se aproximar mais da média das amostras com o aumento do número de amostras.

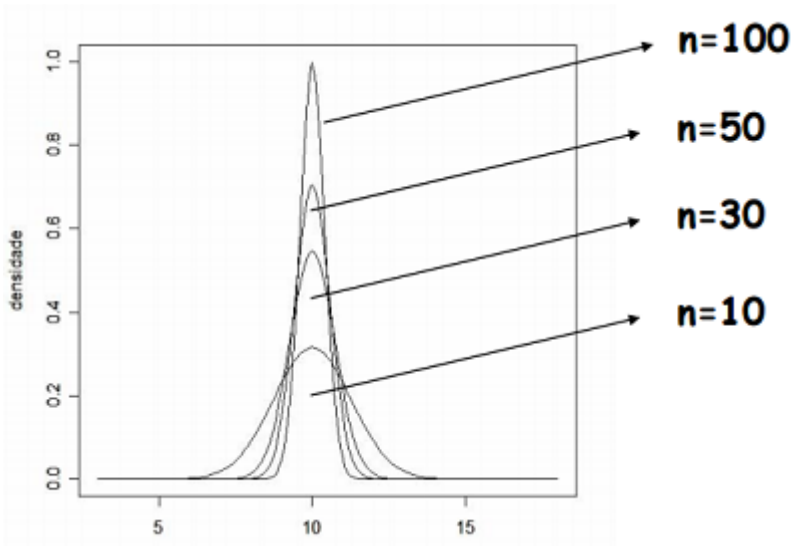


Figura 3.12: Figura demonstrando o efeito de suporte para um número crescente de amostras. O aumento do número de amostras tende a concentrar a função de densidade de probabilidade entorno do valor médio

3.5.2 Teorema do limite Central

Um dos maiores motivos da grande utilização da distribuição normal advém da lei dos grandes números. Certos procedimentos estatísticos são facilitados quando consideramos a convergência de sequências das variáveis com o aumento do número de amostras.

Se uma variável aleatória X puder sr representada pela soma de n variáveis aleatórias independentes, então para um quantidade grande de elementos teremos uma distribuição aproximadamente normal.

Esta propriedade é aceita para qualquer distribuição de probabilidades a priori. Logo temos

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \tag{3.18}$$

Em que S_n é a soma de variáveis aleatórias independentes. Essa distribuição se aproxima da gaussiana padronizada a medida que o número de amostras aumenta.

Teste de hipóteses

Uma hipótese é uma suposição acerca de um dado experimento. Além dos métodos de estimação de parâmetros estamos interessados em tomar decisões que concernem a distribuição de probabilidade daquele valor. Não rejeitar ou rejeitar uma hipótese dependerá da realidade física da estimativa concretizada pela observação das amostras. No entanto, a não rejeição, significa não haver elementos suficientes para averiguar esta possibilidade. A decisão, na verdade, sobre a hipótese não existe. Por uma questão científica, o teste de hipótese é uma ferramenta eliminatória e nunca decisiva, mesmo que em prática se adote a aceitação daquela hipótese para uma dada probabilidade.

Por tratar-se de uma inferência a respeito de uma variável aleatória, o critério de decisão está sempre associado a escolha de um nível de significância α, que corresponde a probabilidade de rejeição da hipótese estabelecida, enquanto $(1-\alpha)$ corresponde ao intervalo de segurança, ou a probabilidade da não rejeição da hipótese nula.

Para um teste de hipóteses geralmente definimos duas alternativas possíveis, uma hipótese nula que queremos anular e uma hipótese alternativa, que guarda outra possibilidade de ocorrência. Considere o problema em uma mineração em que se o valor médio de uma dada impureza no depósito for maior que que 5g/ton o depósito é considerado inviável. Neste caso uma distribuição gaussiana foi ajustada a partir dos dados de amostragem.

$$\begin{cases} H_0 = \mu < 5g/ton \\ H_1 = \mu \geqslant 5g/ton \end{cases}$$

A figura (3.13) demonstra a situação exemplificada.

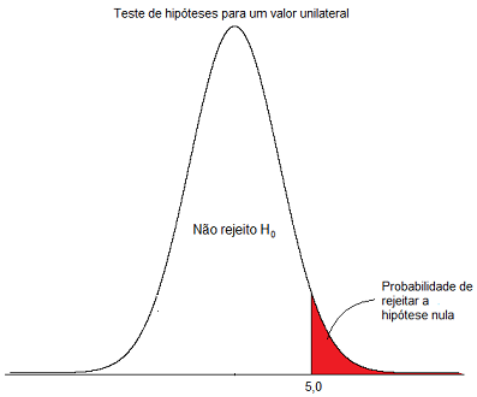


Figura 3.13: Teste de hipóteses para o problema indicado. Área vermelha demonstra a probabilidade de que o valor médio da impureza no depósito exceder a 5g/ton

Este valor demonstrado pela área vermelha também é chamado de valor p e demonstra a probabilidade calculada de se rejeitar a hipótese nula. O valor p é comparado com os limites estabelecidos pelo nível de confiabilidade. Se o valor p for menor que o nível de confiabilidade

opta-se por invalidar a hipótese nula. O procedimento do teste de hipóteses em geral leva aos seguinte procedimentos:

- Definir a hipótese nula e alternativa do problema
- Definir um nível de confiabilidade para o problema
- Definir a estatística de teste (valor médio, variância, proporção)
- Definir a função de densidade de probabilidade para aquela estatística.
- Calcular o valor p, ou a probabilidade de se rejeitar a hipótese nula para a função de densidade de probabilidade do item anterior.
- Comparar o valor p com o nível de significância estabelecido. Se o valor p < nível de confiabilidade, descartar a hipótese nula.

Ajustando uma função de densidade de probabilidade a partir de dados

Após realizado o histograma e a estatística descritiva dos dados pode-se ajustar uma função de densidade de probabilidades. A partir de um teste de hipóteses podemos verificar a aderência do modelo. A ferramenta mais simples para se fazer isso é sem dúvida o gráfico de probabilidade.

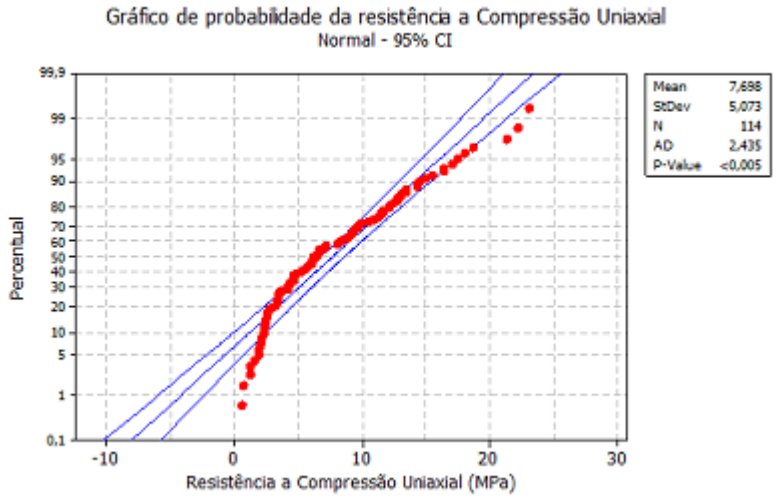


Figura 3.14: Figura demonstrando o gráfico de probabilidade da resistência a compressão uniaxial para uma distribuição de densidade de probabilidade gaussiana com intervalo de confiança de 95%

Note na figura que os pontos estão fora da curva azul em sua maioria. Além disso o valor-p do teste de hipóteses demonstra um valor abaixo de 0.005, que geralmente é o valor admissível para descartar a hipótese nula. Ou seja, neste caso podemos dizer que a variável aleatória não assume comportamento gaussiano.

3.6 Exercícios

Exercise 3.1 Considere o conjunto de amostras com teores de ferro contendo unicamente hematita Fe_2O_3 e sílica SiO_3. Os valores são $(45,69,80,35,56,78)\,\%$. Determine os valores outliers do problema considerando a massa atômica do ferro igual 56g/mol e do oxigênio igual a 16g/mol. Resp.: 80% e 78% ■

Exercise 3.2 Considere o conjunto de amostras com teores $(2.4,5.0,7.6,4.3,2.7,8.9)$ g/ton todos com o mesmo suporte. Encontre o valor da média, da variância, do desvio padrão do

conjunto de amostras. Resp.: $\bar{x} = 5.7$,$s^2 = 5.06$, $s = 2.25$ ■

Exercise 3.3 Um geólogo precisa decidir entre duas metodologias de amostragem para um dado elemento de pesquisa. Entre elas temos a sonda diamantada e o pó de perfuratriz. As incertezas do custo da pesquisa estão diretamente relacionadas com a variabilidade da recuperação, desejando o método com o menor risco associado . Para isso mediu-se a recuperação dos testemunhos e do pó retirado pela máquina. A recuperação dos testemunhos fora de 90% com um desvio padrão de 30%, enquanto a do pó foi de 70% com uma variação de 20%. Deseja-se saber qual método utilizar. Resp.: Pó de perfuratriz « CV ■

4. Estatística bivariada

Na análise de bancos de dados geralmente se torna necessário comparar duas populações diferentes. Em um depósito mineral, por exemplo, podemos ter diversas variáveis presentes. Em alguns casos a relação entre elas pode ser um indício dos fenômenos genéticos de formação das rochas. Em outros casos apenas estamos interessados em como uma informação secundária pode estar relacionada com uma primária de interesse. Seria proveitoso para nós, por exemplo, traçar um modelo que definisse a chance de obter uma amostra com certo teor em contrapartida de outra amostra com o teor de uma variável diferente. Em um depósito vulcanogênico sulfetado podemos estar interessados em prever a quantidade de um elemento metálico a partir do enxofre da rocha encaixante. Enfim, toda a informação que relaciona duas variáveis pode ser descrita pela estatística bivariada.

Diferentemente da estatística univariada, a comparação de histogramas de variáveis diferentes não é uma alternativa interessante sobre o ponto de vista prático. É muito difícil determinar a relação entre duas amostras simplesmente pelas suas proporções individuais. Para isso definimos algumas ferramentas que facilitam ao modelador entender a relação entre duas variáveis distintas visualmente e numericamente.

As seções que se prosseguem mostrarão algumas das ferramentas utilizadas para se caracterizar distribuições bivariadas.

4.1 Gráfico Q-Q plot

O gráfico q-q plot é uma ferramenta para uma primeira análise de diferentes distribuições de variáveis aleatórias. A Figura 4.1 demonstra o gráfico q-q plot das variáveis Cobalto e Cádmio.

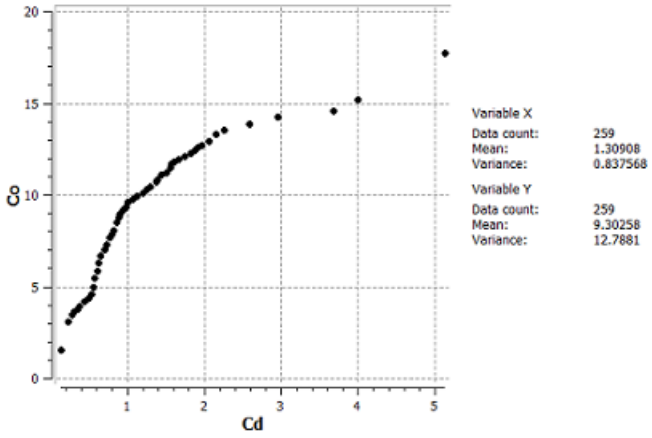


Figura 4.1: Gráfico QQ-Plot de Cobalto e Cádmio. Nota-se uma curvatura característica demonstrando pequena correspondência entre as duas populações. Cada ponto representa o mesmo percentil para cada variável

Neste gráfico são plotados em cada eixo, para um mesmo percentual acumulado das variáveis aleatórias, os valores dos percentis das variáveis. Distribuições idênticas são representadas por uma reta de 45 graus de inclinação no eixo vertical. Distribuições de variáveis com mesma distribuição mas momentos estatísticos diferentes apresentam comportamento linear, mas inclinações diferentes.

4.2 Gráfico p-p plot

Semelhante ao gráfico q-q plot temos o gráfico p-p plot. A figura 4.2 demonstra o gráfico da variável Cobre pela de Cromo.

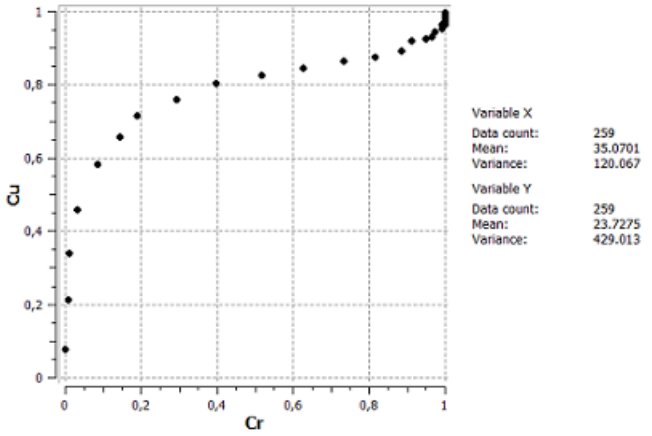


Figura 4.2: Gráfico PP-Plot de Cobre e Cromo. Nota-se uma curvatura característica demonstrando pouca correspondência entre as duas populações. Cada ponto representa o percentual acumulado para o mesmo valor da variável aleatória

A análise do gráfico é feita de forma semelhante ao QQ-plot, no entanto, este gráfico é muito mais sensível à mudança de escala das variáveis. Ele é mais vantajoso quando a ordem de grandeza das variáveis analisadas for semelhante. Neste caso estamos comparando a relação de percentuais acumulados diferentes para o mesmo valor da variável aleatória.

4.3 Gráfico de dispersão

O gráfico de dispersão apresenta dados de duas variáveis dispostos nos eixos cartesianos. Para a utilização do gráfico os dados devem estar colocados. Isso significa que a amostra 1 deve ter a mesma origem da amostra 2, ou o mesmo suporte. Logo só podemos realizar um gráfico de dispersão com vetores de amostras do mesmo tamanho.

Caso a amostragem apresente dados inválidos para uma variável devemos utilizar um filtro para separar apenas os dados colocados. Existem técnicas estatísticas que permitem o tratamento de dados perdidos ou inexistentes, mas nada substitui a amostra em termos de informação sobre o objeto de estudo. A figura (4.3) demonstra um gráfico de dispersão entre a variável cromo e cobalto.

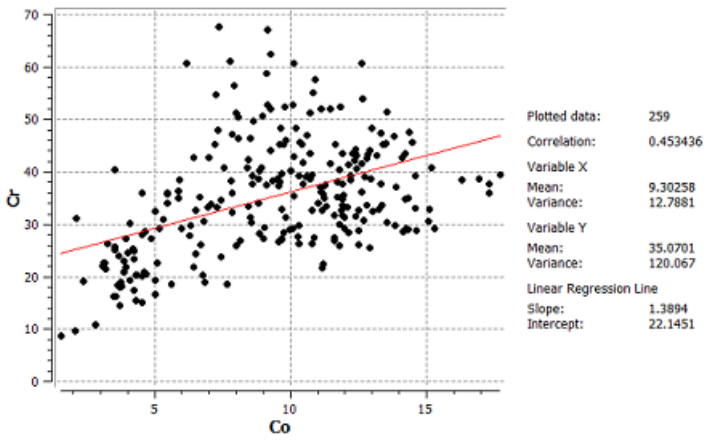


Figura 4.3: Gráfico de dispersão da variável Cromo e Cobalto. Nota-se dependência linear positiva entre as variáveis.

Nota-se pela figura que as variáveis possuem dependência linear positiva entre a variável Cromo e Cobalto. Isso significa que amostras com valor grande de cromo podem apresentar valores grandes de cobalto.

O contrário também pode acontecer, alguns minerais como quartzo e piroxênio são inversamente proporcionais em rochas magmáticas. À medida em que se aumenta o teor de quartzo tende-se a reduzir o teor de piroxênio na amostra de rocha.

A Figura (4.4) demonstra os tipos de correlação lineares possíveis. Em (4.4)a) temos a correlação linear positiva em que o aumento da variável X aumenta o valor de Y, em (4.4)b) temos a correlação linear negativa em que o aumento do valor X tende a diminuir o valor de Y e em (4.4) temos um caso de independência entre as variáveis, tal que o aumento da variável X não altera o valor da variável Y.

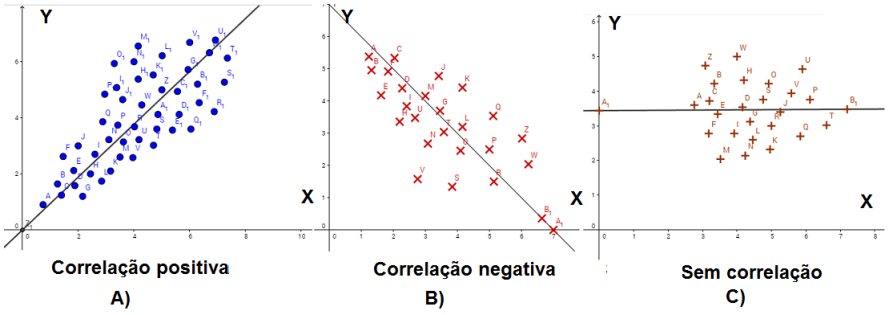


Figura 4.4: Figura demonstrando os tipos de correlação linear possíveis. A) Correlação linear positiva, B) Correlação linear negativa, C) Sem correlação

Os gráficos de dispersão também são uma boa medida para a visualização de valores outliers. A figura (4.5) demonstra a dispersão anterior mas com uma área circulada de pontos que não estão dentro do comportamento linear das variáveis. Neste caso para valores intermediários de Cobalto temos grandes valores de Cromo.

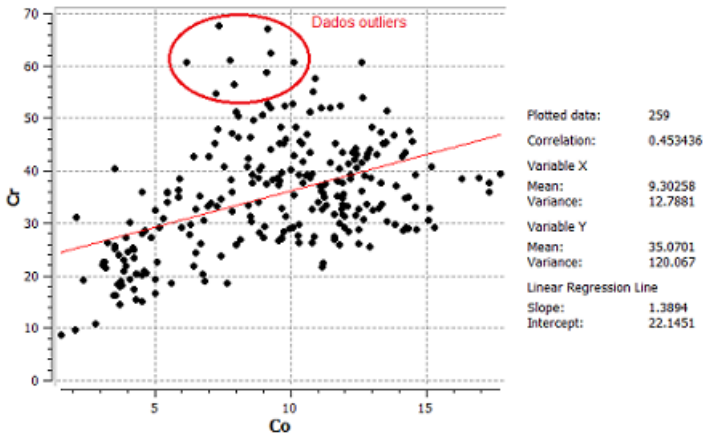


Figura 4.5: Gráfico de dispersão da variável Cromo e Cobalto demonstrando valores outliers. Círculo vermelho indica possíveis valores fora dos padrões das variáveis conjuntas

Muitas vezes um valor outlier em um gráfico bivariado não é demonstrado no tratamento individual das amostras. Muito cuidado deve ser tomado para a retirada de pares anômalos das estatísticas, pois eles podem gerar novos valores discrepantes e não demonstrarem um padrão de maior correlação entre as variáveis.

A figura (4.5) também demonstra uma regressão linear dos dados para a amostra de Cobalto e Cromo. Essa linha em vermelho significa que para cada valor de Cobalto está associado um valor médio de Cromo. Note que a regressão linear é um comportamento médio da variável conjunta e não a sua realização.

4.4 Regressão linear

O modelo de regressão linear simples é aquele em que definimos uma dependência diretamente proporcional entre a variável dependente Y e idependente X. Da mesma forma como demonstrado na krigagem do capítulo 1, o problema da regressão linear pode ser considerado a partir do método

dos mínimos quadrados. Assumindo o modelo determinado pela equação (4.1) temos que

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{4.1}$$

Em que Y_i é o valor médio da variável Y no valor da variável independente X_i para um coeficiente angular β_i, um coeficiente linear β_0 e um erro ε associado ao valor médio.

Isolando o valor do erro da média da variável independente temos que o modelo determinado pela equação (4.2)

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i \tag{4.2}$$

Elevando ao quadrado o erro temos (4.3)

$$\varepsilon_i^2 = (Y_i - \beta_0 - \beta_1 X_i) \tag{4.3}$$

Tomando a soma de todos os erros para cada ponto no gráfico de dispersão temos a equação (4.4):

$$\sum_i \varepsilon_i^2 = \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{4.4}$$

O problema de regressão linear se torna então um problema de otimização ao encontrar o menor somatório dos desvios quadrádicos. A figura (4.6) demonstra graficamente o problema da regressão linear. Neste caso os valores dos coeficientes lineares das retas podem ser encontrados a partir de derivação simples ou por meio de métodos numéricos, como a utilização do método Simplex.

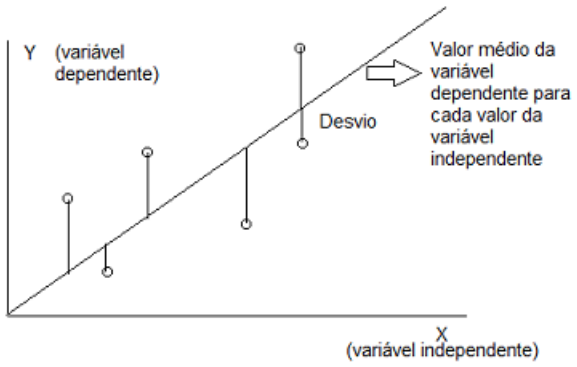


Figura 4.6: Explicação da regressão linear entre a variável independente X e a variável dependente Y. Barras verticais representando o os desvios das amostras com o valor médio.

4.5 Intervalo de segurança para a regressão linear

Ao denotarmos a regressão linear entre duas variáveis é interessante estipular um intervalo de segurança para este valor, lembrando que a média nunca pode estar sozinha de uma outra estatística para que agregue mais informações sobre as amostras. A fórmula (4.5) demonstra como podem ser calculadas as bandas de incerteza da regressão.

$$Y_i \pm t^*_{n-2} s_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}} \tag{4.5}$$

Em que x_i é o valor da variável X a ser estimada a curva da banda, t^*_{n-2} é o valor da distribuição de t-student para um grau de liberdade igual a n-2 e s_y pode ser demonstrado segundo a equação (4.6)

$$s_y = \sqrt{\frac{\sum_i (y_i - Y_i)^2}{n-2}} \tag{4.6}$$

Em que y_i é o valor da coordenada y para um ponto amostral i. Ou seja s_y é o valor do desvio padrão entre os valores amostrais e os valores médios estimados pela regressão.

A figura (4.7) demonstra o intervalo de segurança para o valor regredido.

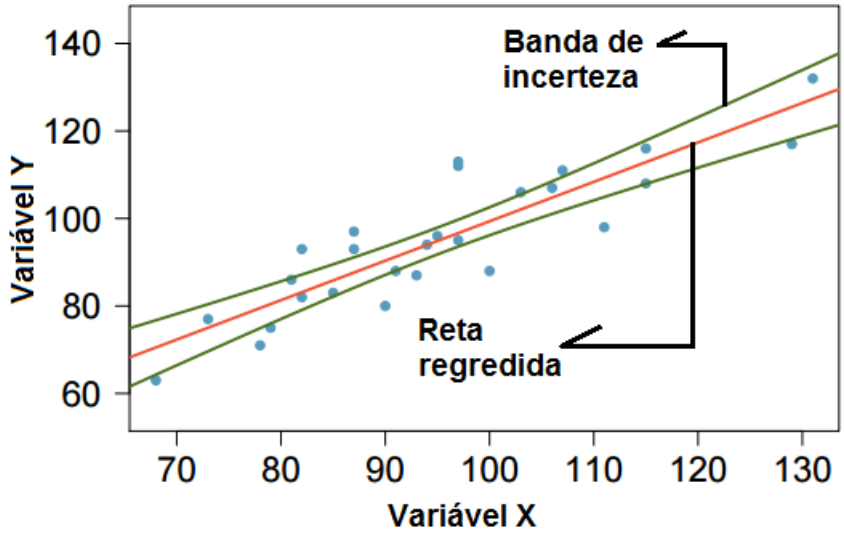


Figura 4.7: Demonstração do intervalo de confiança para a regressão linear. Banda de incerteza adicionada como limite inferior e superior dado pela equação (4.5)

Nota-se que as bandas apesar de acompanharem o valor de regressão linear não são retas, apresentado um maior estreitamento na região mediana da dispersão.

4.6 Regressão linear múltipla

O modelo de regressão linear múltipla é um caso extendido da regressão linear simples para múltiplas variáveis. Neste caso temos um conjunto de n variáveis independentes e n variáveis dependentes, tais que para cada variável dependente temos uma combinação linear de variáveis independentes.

$$Y_i^1 = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \beta_3 X_i^3 ... \beta_n X_i^n + \varepsilon_i \tag{4.7}$$

Isso resulta em uma matriz de valores correlacionados como demonstrado na equação (4.8)

$$\begin{Vmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{Vmatrix} = \begin{Vmatrix} x_1^1 & x_2^1 & .... & x_n^1 \\ x_1^2 & x_2^2 & .... & x_n^2 \\ .... & .... & .... & .... \\ x_1^n & x_1^n & .... & x_1^n \end{Vmatrix} \begin{Vmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{Vmatrix} + \begin{Vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{Vmatrix} \tag{4.8}$$

O erro quadrádico pode então ser determinado por (4.9)

$$\sum_i \varepsilon_i^2 = \sum_i \left(Y_i - \beta_0 - \beta_1 X_i^1 - \beta_2 X_i^2 - ... - \beta_n X i^n\right)^2 \tag{4.9}$$

O resultado dos coeficientes β podem ser encontrados a partir de derivação parcial das equações estabelecidas. A krigagem possui similaridades muito grandes com a regressão linear múltipla, no entanto, possuí restrições de viés durante o processo de otimização realizado.

4.7 Coeficiente de correlação

Vimos no capítulo 1 o conceito de covariância. Esta estatística nada mais é que uma medida da dependência linear entre duas variáveis. Variáveis conjuntas podem apresentar modelos de dependência não lineares como parabólicos, exponenciais ou hiperbólicos. No entanto, cada um destes modelos pode ser transformado em um caso simples de correspondência linear. Por esse motivo torna-se tão importante a regressão linear entre variáveis distintas.

Observe a distribuição de Gates-Gaudin-Schumann comumente utilizado para demonstrar a proporção de diâmetros de partículas de minerais após a cominuição no tratamento de minérios. O modelo é exponencial demonstrado pela equação (4.10)

$$p(d) = \left(\frac{d}{k}\right)^m \tag{4.10}$$

Em que p(d) é a proporção das partículas de diâmetro médio d e k e m são constantes do problema. Tomando o logaritmo desta expressão obtemos a seguinte expressão (4.11)

$$log(p(d)) = m(log(d) - log(k))) \tag{4.11}$$

Em que m e k são as variáveis desconhecidas da distribuição. Esse é um caso simples de linearização de um modelo. Alguns casos mais robustos exigem operações mais complexas, mas a grande maioria dos problemas podem ser resolvidos com linearizações simples.

Uma forma simples de demonstrar o grau de dependência linear entre duas variáveis é utilizando o coeficiente de regressão de Pearson. Ele nada mais é que uma normalização da covariância a partir das variâncias de cada variável aleatória. Definimos o coeficiente de correlacao de Pearson segundo a equação (4.12)

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \tag{4.12}$$

O coeficiente de correlação é uma medida que varia de -1 a 1 sendo o valor negativo igual à uma relação decrescente entre as variáveis, positivo igual uma relação crescente e 0 como independência entre as variáveis aleatórias.

Há também o chamado coeficiente de Rank. Um rank é uma varíavel que associa uma ordem para cada valor da variável aleatória. O menor valor recebe rank igual a 1, o segundo menor recebe um rank igual a 2 e assim sucessivamente até os n valores da variável aleatória. O coeficiente de Rank é então tomado como o coeficiente de correlação entre os ranks de cada variável.

O coeficiente de Rank diferentemente do coeficiente de Pearson é uma medida de tendência dos valores e não da linearidade dos valores em si. Nesse caso estamos interessados em saber apenas se existe uma correlação entre a organização das amostras sem considerar sua ordem de grandeza.

Neste caso a estatística é muito menos influenciada por valores extremos. A tabela (4.1) demonstra suscintamente como realizar o coeficiente de Rank para uma dada variável aleatória.

Tabela 4.1: Valores de Rank para uma variável aleatoria

valor da variável	Rank
50	4
25	1
60	5
30	2
45	3

4.8 Probabilidades condicionais e conjuntas

Muitas vezes não estamos interessados em determinar as probabilidades ou frequências individuais de uma variável aleatória. É interessante, por exemplo, determinar qual é a frequência de um minério e que seu conteúdo tenha um determinado valor de impureza. Ou, por exemplo, qual é a frequência de acidentes e de mal uso de um EPI. Neste caso estamos trabalhando com eventos disjuntos, ou seja a interseção de valores entre variáveis aleatórias. A figura (4.8) demonstra em um diagrama de Venn a interseção formada por um evento disjunto entre duas variáveis.

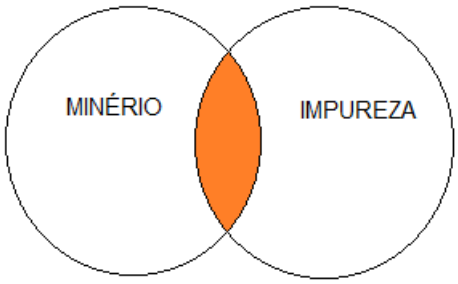


Figura 4.8: Demonstração de eventos disjuntos entre a variável minério e impureza a partir de um diagrama de Venn.Nota-se a área laranja como sendo a interseção dos eventos

Observe a tabela 4.2. Notamos na coluna três o número de vezes que o minério considerado possui uma impureza maior ou igual a 0,005. Neste caso sabemos que há 2 valores em cinco em que isso ocorre. Logo a probabilidade conjunta é P(Minerio) $\bigcap$ P(Impureza $\geq$ 0.005) = 2/5 = 40%

Tabela 4.2: Tabela da relação entre um dado minério e uma impureza

Minério	Impureza	Minério $\bigcap$ (Impureza >=0,005)
Sim	0,005	1
Não	0,007	0
Não	0,008	0
Sim	0,006	1
Sim	0,003	0

Da mesma forma podemos definir a probabilidade condicional como sendo a relação entre a

probabilidade do evento disjunto entre as variáveis aleatórias e a probabilidade da variável aleatória. A probabilidade condicional pode ser demonstrada na figura (4.9) como a relação da área laranja pela área hachurada.

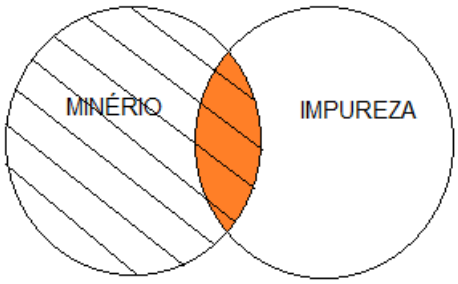


Figura 4.9: Demonstração da probabilidade condicional entre a variável minério e impureza a partir de um diagrama de Venn.Nota-se a área laranja como sendo a interseção dos eventos disjuntos. A probabilidade condicional é a relação entre a área laranja pela área hachurada.

Na tabela 4.2 podemos calcular a probabilidade condicional tal como $P(M = Sim|I \geq 0,005) = \frac{2/5}{3/5} = 67\%$. Ou seja, dado que encontremos um minério, a chance de ele possuir valor acima de impureza acima de 0,005 é de 67%

A figura (4.10) demonstra como pode ser calculado o histograma marginal de uma variável a partir de um gráfico de dispersão. Todos os valores abaixo de 5 da variável Y são selecionados para realizar o histograma.

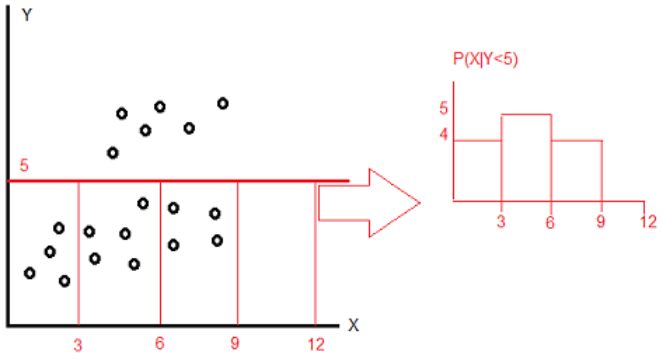


Figura 4.10: Histograma condicional criado a partir de um gráfico de dispersão. Valores abaixo de 5 da variável Y são selecionados para realizar o histograma da direita.

4.9 Teste qui-quadrado para independência entre variáveis

Em alguns casos é importante verificarmos se determinado conjunto de variáveis é independente. No âmbito da geoestatística, variáveis estimadas que são independentes produzem o mesmo resultado que variáveis estimadas em conjunto. Observe a tabela (4.3). A tabela mostra o nível de 4 poços controlados durante 4 dias seguintes. Deseja-se saber se os poços são dependentes e interconectados.

Tabela 4.3: Nivel de quatro poços controlados por dia

	poço 1	poço 2	poço 3	poço 4	soma
dia 1	10	30	20	40	100
dia 2	20	20	25	35	100
dia 3	5	10	20	30	55
dia 4	20	30	30	30	110
soma	55	90	95	135	375

O teste para a verificação de dependência é o qui-quadrado, sendo expresso equação:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \tag{4.13}$$

Em que $E_{i,j}$ são as frequências esperadas para cada uma das observações $O_{i,j}$:

$$E_{i,j} = \frac{\left(\sum_{i=1}^{n} O_{i,j}\right)\left(\sum_{i=j}^{n} O_{i,j}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} O_{i,j}} \tag{4.14}$$

O valor da distribuição qui quadrado deve ser avaliado segundo o grau de liberdade do conjunto de amostras. A ideia de graus de liberdade na estatística está relacionado com o número de amostras menos o número de parâmetros estimados. Neste caso o número de graus de liberdade pode ser calculado como o número de colunas da tabela vezes o número de linhas menos um. Logo o grau de liberdade da tabela deste problema é $4*(4-1) = 12$

Consideramos como hipótese nula no teste qui-quadrado para independência das amostras que as frequências esperadas não são diferentes das frequências observadas, logo as amostras são dependentes. Como hipótese alternativa consideramos que as frequências esperadas são diferentes das observadas, logo as amostras são independentes.

1. Se o valor de χ^2 calculado for $\geq$ ao valor tabelado para um nível de significância, então podemos considerar a hipótese nula rejeitada e as amostras são idependentes
2. Se o valor de χ^2 calculado for $\leq$ ao valor tabelado para um nível de significância, então falhamos em descartar a hipótese nula e podemos considerar as amostras dependentes.

Realizando os cálculos para a tabela acima encontramos um valor de $\chi^2 = 17,88$. O valor da distribuição qui-quadrado para um grau de liberdade igual a 12 e um nível de significância de 5% é de 5,22. Como o valor calculado é maior que o tabelado descartamos a hipótese nula de dependência e a as amostras são independentes.

O teste qui-quadrado é flexível e pode ser utilizado em diversas situações em que se deseja comparar as proporções de diferentes variáveis ou modelos. Pode ser utilizado juntamente com o valor p de gráficos de probabilidade para determinar o ajuste de distribuições.

4.10 Exercícios

Exercise 4.1 Os dados da tabela abaixo representam valores de Au e cobre medidos concumitamente nos mesmos testemunhos de sondagem. Com estes dados, pede-se:

a) Determine a covariância dos dados
b) Determine o coeficiente de correlação.
c) As amostras são dependentes positivamente ou negativamente?
d) Faça um gráfico de regressão linear entre as variáveis ouro e cobre

Au	Cobre
0.012	2.0
0.015	2.02
0.013	1.32
0.070	3.45
0.012	1.02
0.067	2.19
0.090	4.01
0.08	3.67
0.012	1.43
0.011	1.01
0.011	1.05

Exercise 4.2 Os dados da tabela abaixo representam valores de X e Y. Faça um gráfico de dispersão e determine o par de valor outilier para o gráfico.

X	Y
0.729	1.546
0.757	1.683
0.140	0.175
0.575	0.963
0.408	0.726
0.402	1.104
0.616	1.321
0.958	5.02
0.9136	1.873
0.527	0.853
0.470	0.960

5. Técnicas de desagrupamento

Uma amostragem dita representativa é aquela que guarda as mesmas proporções e informações da população. No caso de amostragens espaciais nem sempre podemos resguardar posições espaciais equivalentes entre as amostras. Na mineração temos problemas relacionados com a topografia, áreas de proteção ambiental entre outros motivos que tornam a amostragem impossível em certas áreas. Além disso, por motivos econômicos, é de interesse amostrar áreas que são mais econômicas do que áreas pobres do depósito mineral. Para garantir estatísticas não enviesadas do objeto de estudo utilizamos técnicas de desagrupamento que permitem dar pesos diferenciados para cada amostra durante o cálculo. Note que as técnicas de desagrupamento não devem nunca alterar o valor da amostra considerada, mas apenas a estatística, como por exemplo o valor médio, a variância, a frequência do histograma, etc.

5.1 Estatísticas desagrupadas

Uma estatística desagrupada é aquela em que os valores das amostras recebem um determinado peso associado. A mais comum delas é a média declusterizada, demonstrado na equação (5.1)

$$\overline{Z_d} = \sum_{i=1}^{n} \rho_i z(x_i) \tag{5.1}$$

Em que ρ_i é o peso associado a cada valor de amostra $z(x_i)$. Cada um dos pesos do valor médio de Z são definidos por alguma técnica de desagrupamento como será visto em seções posteriores. Da mesma forma a variância desagrupada pode ser determinada por (5.2)

$$Var_d(Z) = \sum_{i=1}^{n} \rho_i \left(z(x_i) - \overline{Z_d}\right)^2 \tag{5.2}$$

Uma estatística não enviesada é aquela que seu valor esperado tende a convergir para o valor do parâmetro a ser estimado. No caso das estatísticas declusterizadas temos a condição de que a soma dos ponderadores deve ser igual a 1. Isso pode ser provado om a seguinte demonstração

Demonstração. $E(\overline{Z_d}) = E\left(\sum_{i=1}^{n} \rho_i Z(x_i)\right)$
$E(\overline{Z_d}) = \sum_{i=1}^{n} \rho_i E(Z(x_i)) = m$
$E(\overline{Z_d}) = \sum_{i=1}^{n} \rho_i m = m$
$\sum_{i=1}^{n} \rho_i = 1$

■

Essa mesma demonstração ocorre também quando definimos condições para os pesos de krigagem.

O histograma ponderado também é uma ferramenta importante para a análise estatística de dados desagrupados. Neste caso cada frequência de cada classe é alterada segundo a equação (5.3)

$$freq = \frac{\sum_{j=1}^{p} \rho_j}{\sum_{i=1}^{n} \rho_i} \tag{5.3}$$

Em que p é o número de elementos contidos dentro da classe e z_j são os valores contidos em cada classe. O valor n é relativo ao número de todas as amostras consideradas e z_i é o valor de cada amostra.

5.2 Definindo os pesos de desagrupamento

5.2.1 Método dos polígonos de influência

O primeiro método para encontrar pesos para cada uma das amostras é conhecido como polígonos de influência ou de Thyssen. O objetivo é encontrar uma área convexa para cada uma das amostras contendo a região de influência. Para o caso tridimensional os polígonos se transformam em um poliedro de influência, sendo muito mais complexo a determinação matemática do volume de influência, tornando da metodologia muito mais relevante e desejável para casos bidimensionais. Computacionalmente os polígonos ou poliedros de influência podem ser determinados dividindo o espaço entre as amostras por células, sendo a definição da área ou do volume condicionada ao tamanho da célula, ou seja, quanto menor for a dimensão, maior será a resolução das regiões de influência.

A figura (5.1) demonstra o primeiro passo para se determinar o polígono de Thyssen manualmente. Divide-se as amostras em dois subespaços determinados pela equidistância entre as amostras.

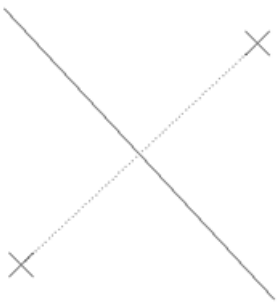


Figura 5.1: Demostração da divisão em dois subespaços entre as amostras. Distância equivalente entre as duas amostras

A figura (5.2) demonstra a divisão de todos os subespaços entre todas as amostras mais próximas da amostra considerada. O polígono convexo é formado pela interseção de todas as linhas mais próximas da amostra em que se deseja calcular o polígono.

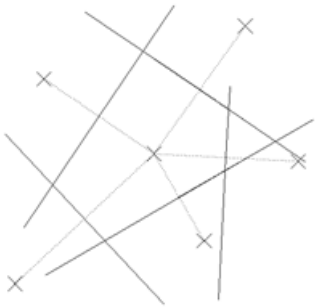


Figura 5.2: Determinação do polígono convexo entre todas as amostras mais proximais da amostra considerada

Cada vértice do polígono de Thyssen também pode ser determinado como o centro do círculo determinado pela amostra e três pontos mais próximos. A figura (5.3) demonstra esta relação geométrica.

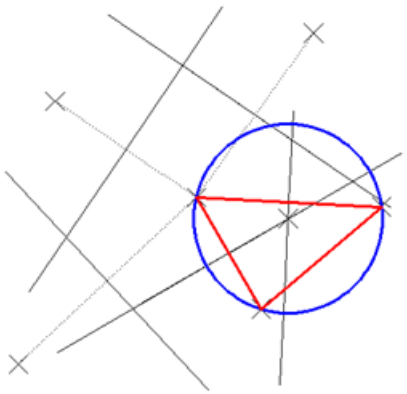


Figura 5.3: Determinação do vértice do polígono de Thyssen pelo círculo determinado pela amostra considerada e duas amostras mais proximais

Cada amostra receberá então um peso igual a sua área de influência dividido pela área total de todas as amostras. A equação (5.4) demonstra o cálculo de cada peso para cada amostra i utilizando os polígonos de influência.

$$\rho_i = \frac{A_i}{\sum_{i=1}^{n} A_i} \tag{5.4}$$

Em que A_i é a área ou volume de influência de cada amostra.

5.2.2 Método das células móveis

Outra forma de se determinar os ponderadores utilizados para o desagrupamento das amostras são as células móveis. O espaço entre as amostras é dividido em uma malha regular com n blocos. A figura (5.4) demonstra um conjunto de amostras dividido em um grid de quatro blocos.

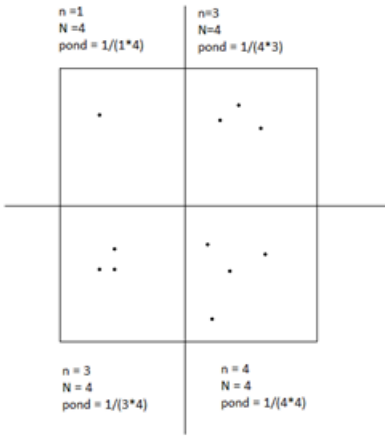


Figura 5.4: Demonstração do método das células moveis dividindo as amostras em quatro células. Cada amostra receberá um peso como demonstrado na figura

Cada peso de cada amostra é dado pela relação (5.5)

$$\rho_i = \frac{1}{nN} \tag{5.5}$$

Tal que n é o número de amostras contidas na célula e N é o número de células. Neste caso todas as amostras que estão contidas dentro de uma célula recebem valores iguais de peso, sendo uma desvantagem do método. O processo de desagrupamento consiste em achar o tamanho de célula que minimiza a estatística desagrupada, e para isso é necessário um processo interativo ao qual são escolhidos vários tamanhos até se determinar aquele que melhor se adequa a situação geométrica das amostras. A figura (5.5) demonstra como podemos encontrar a média desagrupada a partir de um gráfico em que variamos o tamanho das células para encontrarmos o mínimo da função. Os pesos que se adequam a menor média desclusterizada são os escolhidos para o desagrupamento dos dados.

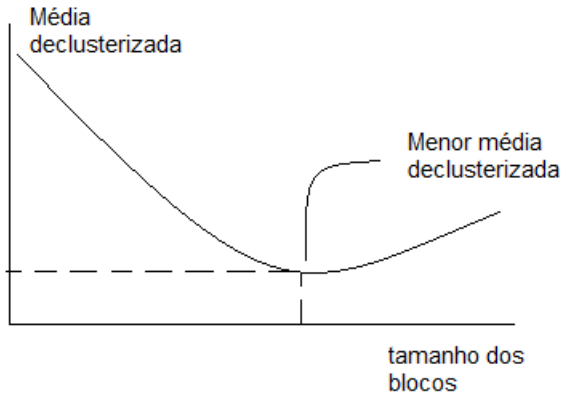


Figura 5.5: Demonstração do método das células moveis dividindo as amostras em quatro células. Cada amostra receberá um peso como demonstrado na figura

5.2.3 Regularização de amostras

Uma das condições dos métodos de estimativa e simulação utilizados na geoestatística é que as amostras devem possuir o mesmo suporte. No entanto isso nem sempre é possível na prática. Na

avaliação de depósitos minerais, por exemplo, os testemunhos de sondagem apresentam tamanhos diferenciados e recuperações diferenciadas. Para isso devemos regularizar as amostras em um tamanho único e atribuir o valor médio para os testemunhos. Em casos que o suporte do volume estimado ou simulado for muito maior do que o suporte das amostras a regularização praticamente não possui efeito.

O primeiro passo para efetuar a regularização é estipular um tamanho para ser regularizado. Isso pode ser feito encontrando a moda dos valores dos tamanhos das amostras. Em seguida definimos a interseção das amostras com o suporte regularizado. A figura (5.6) demonstra uma situação em que temos um minério M1 e M2. O suporte regularizado contém parte de M1 e M2.

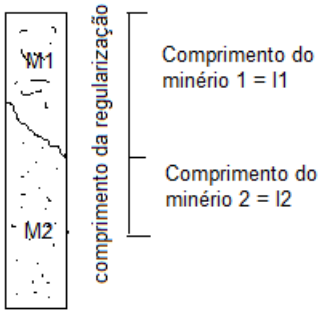


Figura 5.6: Exemplo da regularização de um testemunho com dois minérios M1 e M2. Dentro do comprimento regularizado temos a partipação de um comprimento do minério 1 = l1 e um comprimento do minério 2 = l2

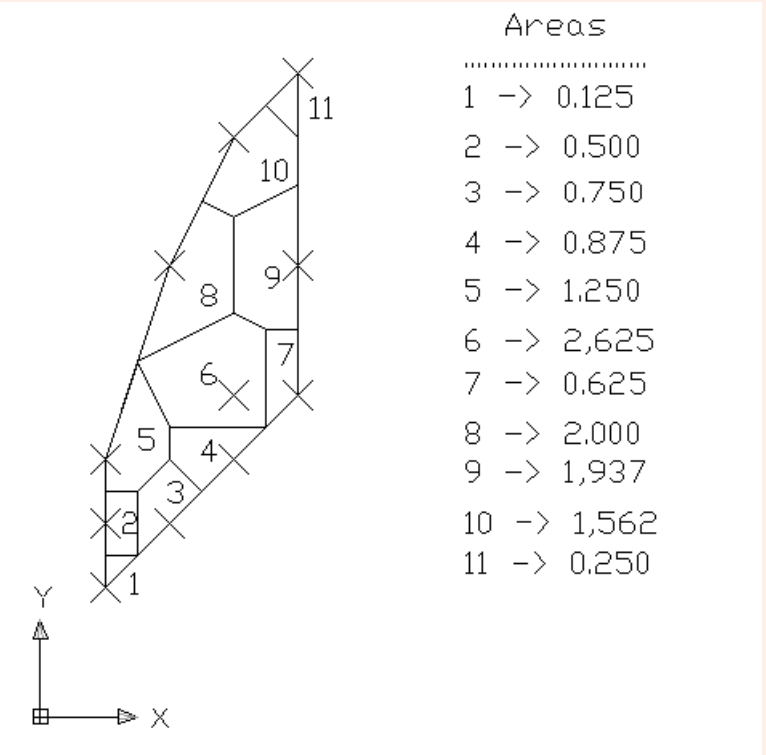
Para definir, por exemplo, o teor médio do suporte regularizado, temos que realizar os cálculos segundo a equação (5.6)

$$\bar{z} = \frac{l_1 z_1 + l_2 z_2}{l_1 + l_2} \tag{5.6}$$

Sendo z_1 e z_2 os teores médios de cada amostra e l_1 e l_2 os seus comprimentos. Na verdade apenas definimos a regularização dos teores pelo comprimento porque a seção do testemunho é a mesma para cada uma das amostras. Lidar com variáveis aleatórias regionalizadas é antes de tudo aprender a conhecer os efeitos de escala para o problema tratado. Como a dimensão do comprimento de furos de sondagem é muito maior que sua seção horizontal é considerável tratar o problema em apenas uma dimensão.

Exercise 5.1 Os dados da tabela abaixo representam um conjunto de amostras bidimensionais, em que x e y representam respectivamente as coordenadas cartesianas nos eixos das abicissas e das ordenadas. Para a configuração geométrica abaixo, determine os polígonos de Thyssen, e consequentemente os ponderadores para cada uma das amostras. (Obs.: Feche os polígonos no limite exterior da região das amostras ligando diretamente as amostras)

x	y	z
1	2	1.09
1	3	0.50
1	4	2.01
2	3	2.04
2	7	7.90
3	4	3.05
3	5	2.02
3	9	3.04
4	5	2.01
4	7	2.01
4	10	3.07



Exercise 5.2 Para os dados do exercício anterior encontre a média declusterizada e a variância declusterizada dos dados.

6. Continuidade Espacial

Até os capítulos anteriores deste livro estávamos preocupados em realizar estatísticas diretamente com os dados. Inicialmente propomos o conceito de continuidade espacial que define o limiar entre as técnicas de estatística e de geoestatística. Como descrito no capítulo um no tópico de krigagem, o modelo de covariâncias é que define o processo de estimativa. Este capítulo apresenta uma breve revisão teórica sobre variografia e os conceitos relevantes para investigar o comportamento espacial do fenômeno geológico. Toda a conceituação teórica envolvida neste capítulo será de relevância para o entendimento dos capítulos posteriores.

6.1 Definição de continuidade espacial e variografia

A continuidade espacial define o comportamento médio da variável regionalizada para direções determinadas. Ela é uma propriedade intrínseca do fenômeno estudado e caracteriza sua organização espacial.

A variografia utiliza de funções estatísticas para reconhecer a continuidade espacial da variável aleatória. São formulações bi-pontuais que requerem a disposição espacial das amostras, conjugando sempre pares de valores para um dado espaçamento. A Figura 6.1 demonstra o posicionamento de amostras dentro de um domínio D e a diferença vetorial de espaçamento h.

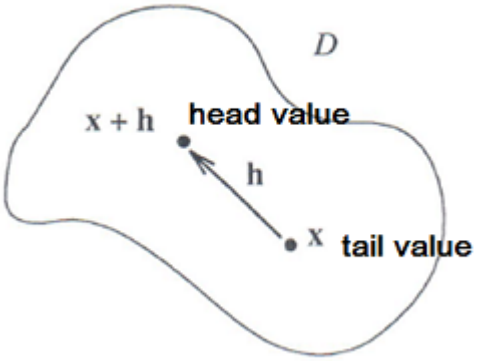


Figura 6.1: Notação geoestatística para a definição de um lag h em uma direção dentro de um domínio D de uma amostra com suporte x. Dois pares de pontos, um considerado head value, ou ponta do vetor e outro considerado tail value, ou início do vetor.

As funções de continuidade espacial medem comportamentos diferentes da variável aleatória: ou elas determinam a similaridade dos valores, ou determinam suas dissimilaridades . Essas propriedades são análogas às projeções vetoriais. A Figura 6.2 a demonstra a dissimilaridade como uma diferença de dois vetores, um vetor considerando os dados in situ e outro com os dados deslocados em um direção h. A diferença entre os vetores é analogamente comparada à função variograma em que são calculadas as médias das diferenças entre dados. Quanto maior for a discrepância entre os dados, maior será o vetor da diferença entre as amostras. A Figura 6.2 b também demonstra a similaridade dos dados como uma projeção vetorial, em que um conjunto mais similar apresenta maiores projeções. Nesse caso a similaridade analogamente comparada com o covariograma pode ser comparada com uma projeção de um vetor sobre outro ou com o produto escalar entre o vetor Z(x+h) e o vetor Z(x).

6.2 Dependência espacial

Para os fenômenos geológicos, é de se esperar que as amostras mais próximas apresentem maior similaridade de valores amostrados. A Figura 6.3 é uma representação gráfica desta característica comentada em um gráfico h-scatter para valores de lag crescentes. Quanto maior for a distância entre as amostras, mais estes pares de valores são dissimilares ou descorrelacionados.

6.3 Hipótese de estacionaridade

As funções de continuidade espacial requerem que o modelo ajustado não seja afetado pela translação. A hipótese estacionaridade também pode ser denominada de invariância de translação. Alguns tipos de funções exigem hipóteses mais fortes, tal como a de estacionaridade segunda ordem. Esta admite que todas as distribuições das variáveis aleatórias no espaço possuam médias e variância iguais. Um exemplo da necessidade da estacionaridade de segunda ordem é a portabilidade das funções variograma e covariograma. As funções variograma e covariograma podem ser relacionadas quando estabelecido um regime estacionário de segunda ordem, resultando na Equação 6.1:

$$\gamma(h) = C(0) - C(h) \tag{6.1}$$

Em que $\gamma(h)$ é o semi-variograma, $C(0)$ é a variância à priori dos dados e $C(h)$ o valor do covariograma. No entanto, a presença de tendência nos dados mostra que o covariograma e o

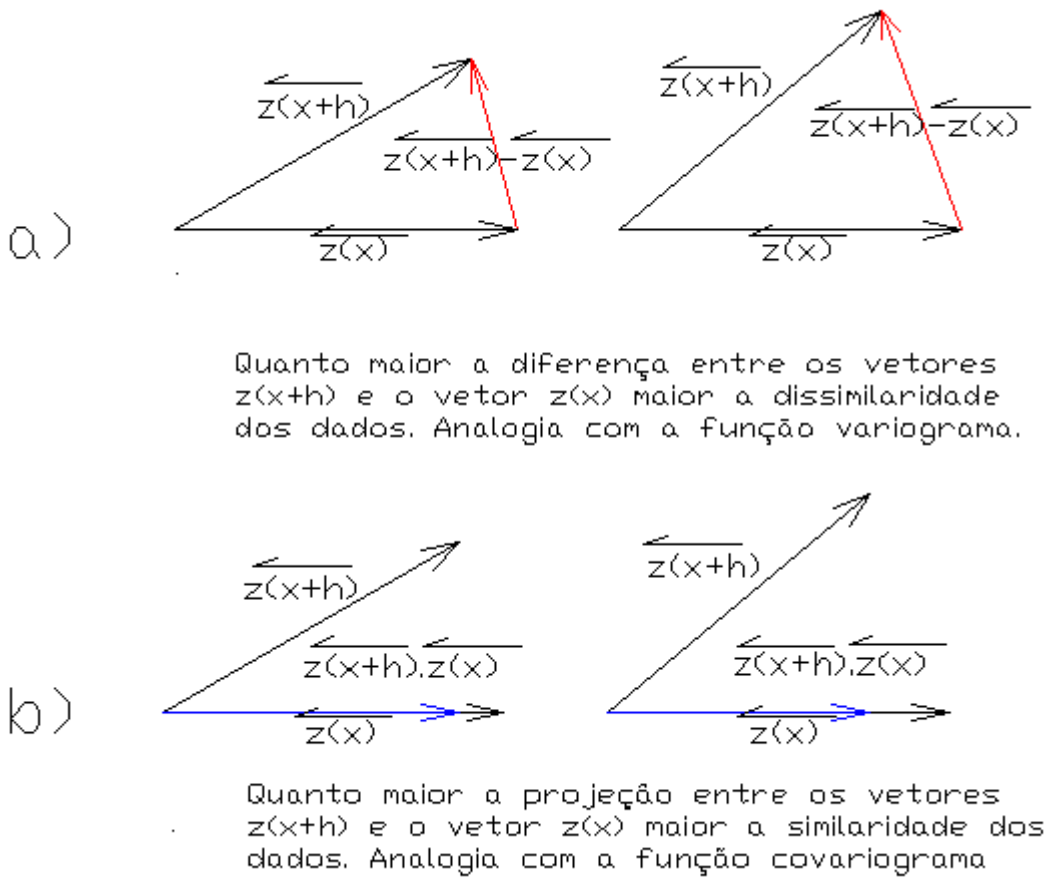


Figura 6.2: Funções de continuidade espacial como uma interpretação de vetores. a) Variograma como uma diferença de vetores b) Covariograma como uma projeção de vetores.

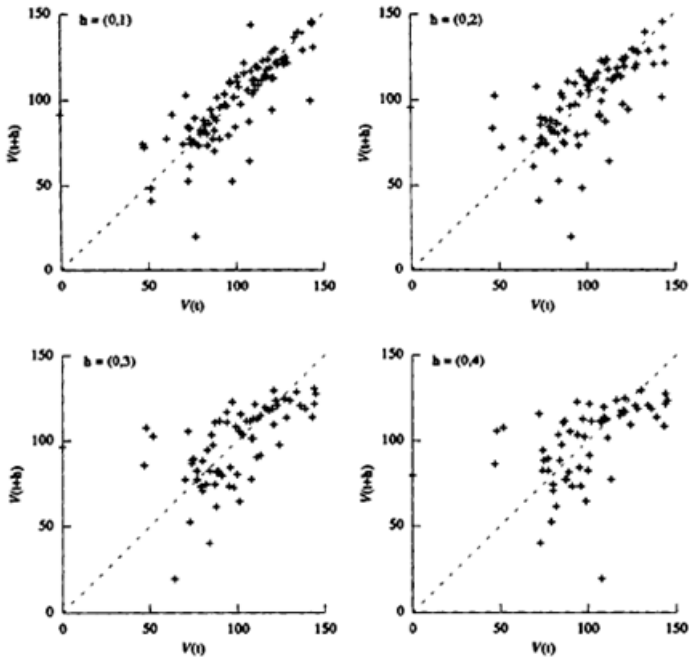


Figura 6.3: H-scatterplots de uma variável para lags de 0,1m , 0,2m, 0,3m e 0,4m. Aumento da descorrelação de acordo com o aumento do comprimento dos lags.

variograma não se convertem em uma relação direta, pois a média da variável aleatória não é constante no domínio e por isso não implica em uma hipótese de estacionaridade de segunda ordem.

6.4 Funções experimentais de continuidade espacial

6.4.1 Efeito dos dados sobre os valores experimentais

As funções clássicas de continuidade espacial são afetadas por valores extremos, esparcidade dos dados e valores clusterizados o que levou à investigação de funções de estimativas robustas. Leva-se em consideração que a continuidade espacial é uma propriedade do domínio e não das amostras. No entanto, pela escassez de informação, ela é inferida a partir de uma quantidade limitada de dados. Se as observações não cobrirem as dimensões do objeto de estudo, devido a um número pequeno de amostras com dados esparsos ou agrupados, a estimativa pode não representar a continuidade do fenômeno. Pode-se demonstrar a conexão entre o variograma experimental e a amostragem, tal que as amostras devem respeitar as seguintes definições:

1. As amostras devem estar contidas na mineralização do depósito.
2. Os corpos de minério devem ser tratados de forma diferenciada.
3. Todas as amostras devem ter o mesmo suporte.

Um número crescente de amostragens pode nem sempre resultar em um benefício da informação sobre a continuidade espacial. Como as funções de continuidade espacial são valores médios de uma gama de pares de amostras no espaço, e os valores médios tendem a suavizar o efeito das informações, o acréscimo de pares de informações redundantes pode não alterar as funções de continuidade espacial. O reconhecimento da continuidade exige uma estratégia de amostragem adequada e depende complexidade do depósito mineral. Corpos com baixa continuidade espacial podem ser analisados com malhas adensadas em contrapartida de depósitos com alta continuidade espacial que podem ser analisados com malhas menos adensadas.

6.4.2 Funções de continuidade espacial mais comuns

Matheron o criador da geoestatística desenvolveu as principais funções de estimativa da continuidade espacial para entender o comportamento das variáveis aleatórias regionalizadas. Duas destas são consideradas as mais tradicionais definidas inicialmente pelo autor. O covariograma e o variograma medem respectivamente a similaridade e dissimilaridade dos dados. Define-se a função covariograma pela Equação 6.2:

$$C(h) = E\left[(Z(x+h) - m(x+h))\,(Z(x) - m(x))\right] \tag{6.2}$$

Em que $Z(x)$ é o valor da variável aleatória no suporte i, $Z(x+h)$ é o valor da variável aleatória transladada por um vetor h e $m(x+h)$ o valor médio da variável transladada. Sob a hipótese de estacionaridade de segunda ordem os valores da média são constantes tal que $m(x+h) = m_i$ e a Equação 6.2 pode ser traduzida pela Equação 6.3 por uma transformação algébrica:

$$C(h) = E\left(Z(x+h)Z(x)\right) - m^2 \tag{6.3}$$

A função variograma pode ser representada pela Equação 6.4:

$$2\gamma(h) = E\left[(Z(x+h) - Z(x))^2\right] \tag{6.4}$$

Em que $\gamma(h)$ é também denominado de semi variograma. Na literatura, é comum a utilização ambígua dos termos, referindo-se ao valor de semi variograma como a função variograma. A função semi variograma pode ser representada como uma distância da dispersão de pontos em relação à reta Y=X, em um gráfico h-scatterplot. A Figura 6.4 é uma demonstração gráfica da interpretação do semivariograma como um valor médio das distâncias das variáveis Zu e Zu+h, com esses valores em x e y e com a reta de correlação máxima. A demonstração da Figura 6.4 em termos matemáticos está descrito na Equação 6.5:

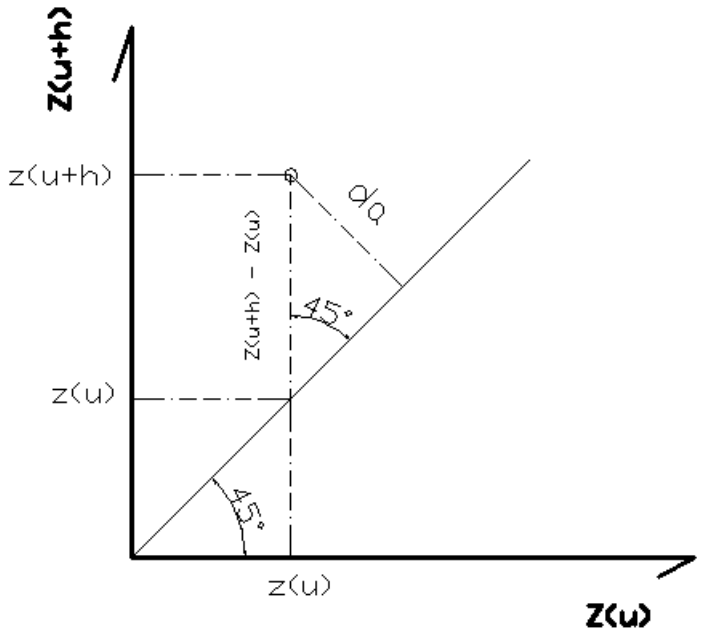


Figura 6.4: Interpretação geométrica do semi-variograma como sendo a distância de um ponto em relação a reta x=y.

$$\frac{1}{n}\sum_{i=1}^{n} da^2 = \frac{1}{n}\sum_{i=1}^{n} sen^2(45^\circ)\,(z(x)-z(x+h))^2 = \frac{1}{2n}\sum_{i=1}^{n}(z(x)-z(x+h))^2 \tag{6.5}$$

6.4.3 Outras funções experimentais

Várias são as funções de continuidade espacial utilizadas na bibliografia, principalmente as mais clássicas desenvolvidas por Matheron. No entanto, a busca de estimativas cada vez menos sensíveis aos valores extremos levou ao desenvolvimento de modelos robustos. Entre eles podemos citar o variograma relativo e o pairwise. Há também uma classe de diferentes tipos de estimativas para a função variograma, utilizando uma série de médias com limites aparados e dentre elas a própria mediana para poucos valores de pares.

Modelos mais robustos de variograma estão desenvolvidos ao longo da bibliografia. A Equação 6.6 demonstra a relação determinada pelos autores, que garante menores efeitos de valores extremos ao contrário do variograma tradicional proposto por Matheron, também denominada de Rodograma:

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n} |Z_i - Z_{i+h}|^{\frac{1}{2}} \tag{6.6}$$

A Tabela 6.1 é uma representação das principais estimativas de funções de continuidade espacial. Alguns tipos tem maior recorrência que as demais como o variograma tradicional e a covariância, propostos inicialmente por Matheron:

Em que m_i e $m(x+h)$ são os valores médios determinados no início e na ponta do vetor ($E(Z(x))$ e $E(Z(x+h))$) e $\sigma(x)^2$ e $\sigma(x+h)^2$ são as variâncias no início e na ponta do vetor. No caso de estacionaridade de segunda ordem $m_i = m_{i+h}$ e $\sigma_i^2 = \sigma_{i+h}^2$, no entanto as funções experimentais calculam à priori estes valores de acordo com as distribuições amostrais do tail e do head para cada diferença de lag. A obtenção dos valores experimentais das funções de continuidade espacial é a etapa inicial, que é seguida pela modelagem variográfica e pelas etapas posteriores de estimativa e simulações.

Tabela 6.1: Funções de continuidade espaciais experimentais

Função experimental	Equação
Semi-variograma	$\sum_{i=0}^{n} \frac{(Z_i - Z_{i+h})^2}{2n}$
Covariograma	$\frac{1}{n}\sum_{i=0}^{n}(Z_i - m_i).(Z_{i+h} - m_{i+h})$
Correlograma	$\frac{1}{n}\sum_{i=0}^{n} \frac{(Z_i - m_i).(Z_{i+h} - m_{i+h})}{\sigma_i \sigma_{i+h}}$
Pair-Wise	$\frac{1}{n}\sum_{i=0}^{n} \frac{(Z_i - Z_{i+h})^2}{\left(\frac{Z_i + Z_{i+h}}{2}\right)^2}$
Madograma	$\frac{1}{n}\sum_{i=0}^{n} \lvert Z_i - Z_{i+h}\rvert$
Variograma Relativo	$\frac{1}{n}\sum_{i=0}^{n} \frac{(Z_i - Z_{i+h})^2}{\left(\frac{m_i + m_{i+h}}{2}\right)^2}$

O cálculo dos valores experimentais é realizado segundo uma direção vetorial. A Figura 6.5 demonstra os valores de amostras aceitáveis como pares de pontos permissíveis. Para um lag unitário, somente os valores adjacentes no grid estão disponíveis. A direção escolhida para o cálculo do variograma experimental é leste-oeste.

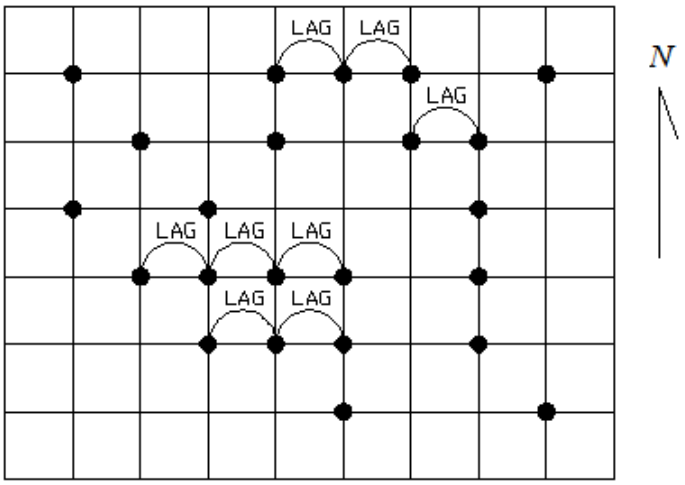


Figura 6.5: Cálculo de variogramas experimentais segundo um lag unitário na direção Leste-Oeste.

O mesmo pode ser representado na Figura 6.6, em que os valores disponíveis como pares para o cálculo são efetuados em dois nós do grid consecutivos.

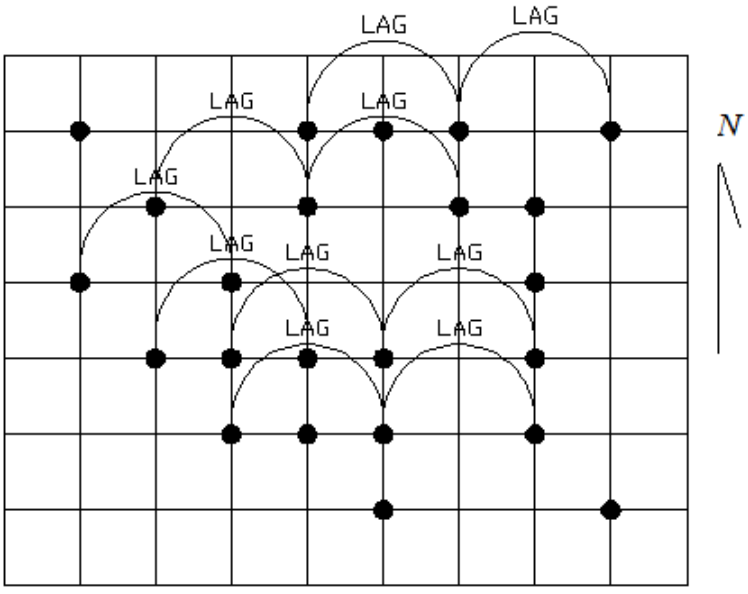


Figura 6.6: Cálculo de variogramas experimentais segundo o dobro do lag na direção leste-oeste.

Os valores calculados do variograma nas Figuras 6.5 e 6.6 estão demonstrados em um gráfico na Figura 6.7 de forma ilustrativa como dois pontos consecutivos, P1 e P2 ligados por um modelo

hipotético como forma ilustrativa. Cada diferença de lag representará um valor de variograma associado.

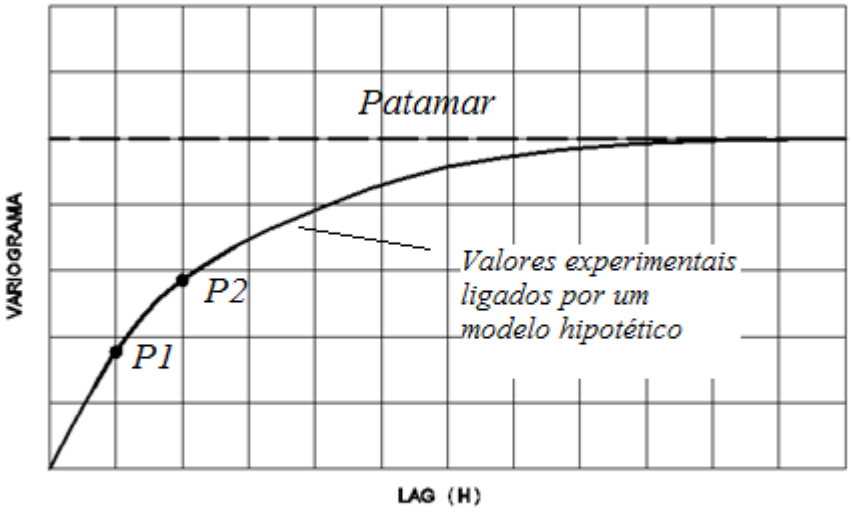


Figura 6.7: Função variograma experimental para os cálculos nas Figuras 9 e 10 e ajuste em um modelo hipotético. P1 e P2 representam os pontos para um lag unitário e o dobro do lag.

6.4.4 Parâmetros de busca

Nas Figuras 6.5 e 6.6, a direção escolhida para o cálculo da função variograma permite medidas regulares. No entanto, a maioria dos casos relacionados à mineração é caracterizada por disposições irregulares das amostras. Neste caso, o variograma direcional não é mais calculado em uma direção absoluta, mas apresenta uma região de incerteza no alinhamento das amostras. A Figura 6.8 é uma representação da busca de pares irregularmente espaçados.

Para o problema bidimensional, são consideradas 3 variáveis geométricas de incerteza e o lag do vetor propriamente dito. As geometrias variáveis são:

1. Tolerância angular = Desvio angular da direção nos lags de menor tamanho.
2. Banda = Desvio lateral da busca.
3. Tolerância linear = Desvio longitudinal da busca.

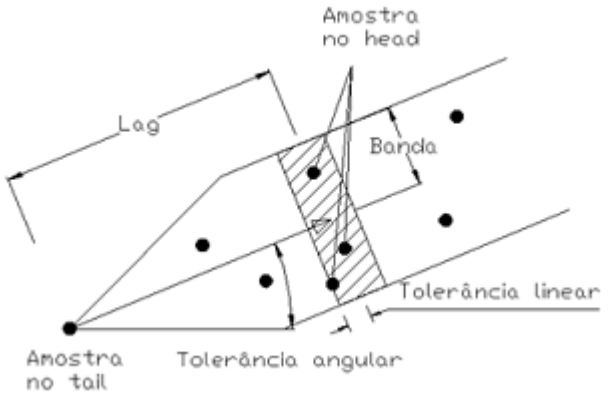


Figura 6.8: Busca de pontos da função de continuidade para amostras irregularmente espaçadas.

No caso tridimensional, a busca de pares pode se realizar de duas formas diferentes, sob uma perspectiva elíptica ou prismática. A Figura 6.9 é uma representação da busca de pares nas duas formas geométricas.

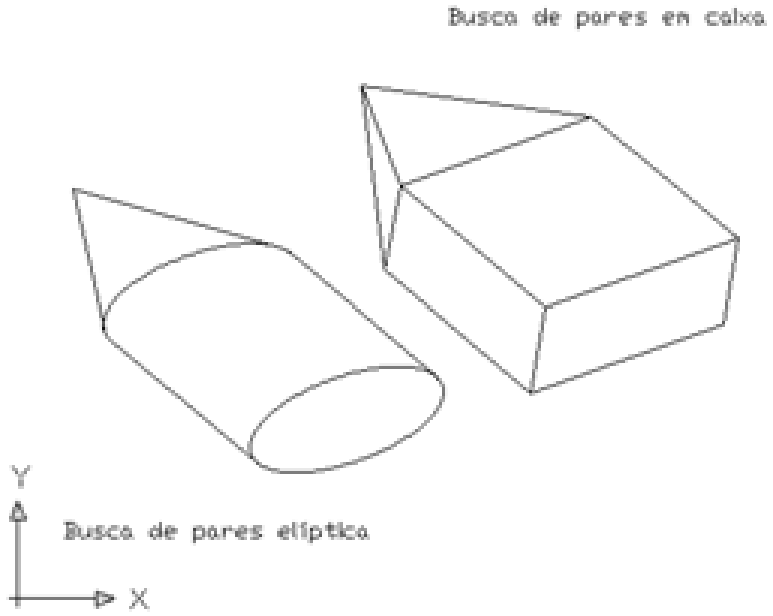


Figura 6.9: Busca de pares de pontos em uma direção tridimensional. Busca de pares elíptica utilizada no SGeMS e em caixa utilizada no GSLib.

A procura pela metodologia em caixa é utilizada no software Gslib, em que são estipuladas não somente a tolerância horizontal e a banda horizontal, como também a tolerância vertical e a banda vertical. A alternativa de caixa é preferencial à busca de pontos cilíndrica, pois em casos onde depósitos minerais apresentem estratigrafia característica, as bandas verticais e horizontais podem ser utilizadas para delimitação de amostras pertencentes ao mesmo nível de formação geológica.

6.5 Modelagem de funções de continuidade espacial

6.5.1 Modelos de variogramas permissíveis

Após a estimativa dos valores experimentais, a análise variográfica procede com a modelagem de funções permissíveis e estabelecimento de um modelo simplificado de regionalização. Um modelo permissível de variograma deve possuir as seguintes características:

1. O modelo deve ser uma função par $\gamma(h) = \gamma(-h)$.
2. O modelo deve ser uma função positiva definida tal que qualquer combinação linear dos seus valores deve ser maior ou igual a zero, como demonstrado na Equação 6.7.

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\lambda_j\gamma(x_i - x_j) \geq 0 \tag{6.7}$$

   Em que λ_i é uma constante de proporcionalidade e x_i e x_j são as diferenças das amostras em um suporte i e j qualquer.
3. Modelo deve ser limitado por um valor limite, geralmente caracterizado como a variância à priori do fenômeno.

6.5.2 Parâmetros das funções de continuidade

O conjunto de variogramas transitivos, ou seja, que apresentam um patamar possuem parâmetros característicos. A Figura 6.10 é uma representação de um modelo de variograma. Os parâmetros da função são:

1. Efeito pepita: Caracteriza a dispersão dos valores para um lag imediatamente maior que zero. O Efeito pepita representa, além da variabilidade de escala, os erros associados à amostragem.
2. Range ou alcance: Máxima distância de influência da correlação. A partir do range não mais existe correlação entre os pares de valores da variável aleatória e estes podem ser ditos independentes.
3. Patamar: O patamar representa o estabelecimento da máxima dispersão admissível. Nas funções em que a similaridade é a propriedade caracterizada, o patamar assume valor nulo tal como na Figura 6.11. A melhor estimativa para o patamar pode não ser a variância das amostras, mas deve-se considerar o posicionamento espacial destas pela utilização da variância de dispersão, da declusterização dos pesos, além do tratamento de valores extremos.

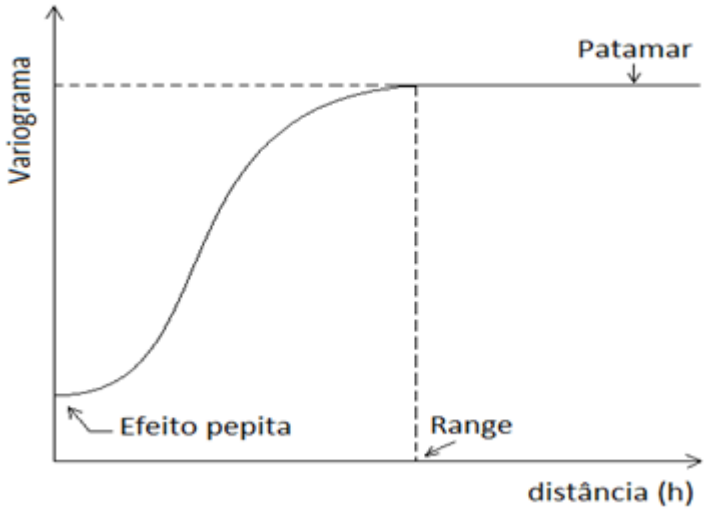


Figura 6.10: Parâmetros do variograma.

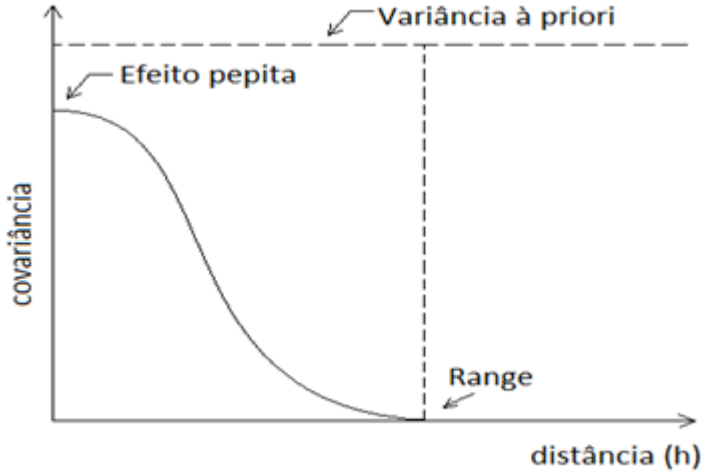


Figura 6.11: Parâmetros da função covariograma.

6.5.3 Modelos de continuidade espacial mais comuns

Dentre os modelos de covariância mais comuns podemos citar:

Efeito de pepita puro: Considera o efeito de dispersão puro. Não representa nenhuma conectividade dos dados e a probabilidade de ocorrência de um determinado valor é caracterizada por uma distribuição uniforme. O efeito pepita pode ser caracterizado como uma percepção não linear do fenômeno em uma escala considerada. A Equação 6.8 demonstra o modelo de variograma com efeito de pepita puro.

$$\gamma(h) = \begin{cases} 0 & ,h = 0 \\ 1 & ,\text{ao contrário} \end{cases} \tag{6.8}$$

Modelo exponencial: O modelo representa o valor de variabilidade com decaimento exponencial. Apresenta-se assíntota no patamar e o range é caracterizado por um valor prático que ocupa 95% da variância a priori quando h = 3a, sendo "a" o alcance prático. A Equação 6.9 demonstra o modelo de variograma exponencial.

$$\gamma(h) = 1 - \exp^{\frac{-h}{a}} \tag{6.9}$$

Modelo Gaussiano: O modelo representa o valor de variabilidade de decrescimento exponencial quadrático. Dentre as funções, é a que apresenta maior suavização próxima da origem. Apresenta também um range prático tal que $h = a\sqrt{3}$. A Equação 6.10 demonstra o modelo de variograma gaussiano.

$$\gamma(h) = 1 - \exp^{-\frac{h^2}{a^2}} \tag{6.10}$$

Modelo Esférico: A Equação 6.11 é a representação de um modelo esférico. Apesar de constituir uma função de terceira ordem, que feriria os princípios de positiva definida, o modelo esférico é limitado pelo alcance da função, e a partir daquele valor é substituído pelo patamar.

$$\gamma(h) = \begin{cases} \left(\frac{3h}{2a} - \frac{h^3}{2a^3}\right) & ,h < a \\ 1 & ,h >= a \end{cases} \tag{6.11}$$

Como todos os modelos prescritos são permissíveis então qualquer combinação destes também resulta em um modelo permissível.

6.5.4 Anisotropia

A anisotropia é a mudança de comportamento das propriedades do variograma por rotação. Os fenômenos geológicos podem permitir a gênese diferenciada dos litotipos à partir de controles e enriquecimentos em sentidos distintos. Dois casos são recorrentes na literatura e envolvem a forma geométrica e zonal.

O caso geométrico delimita alcances diferentes para um mesmo patamar. A anisotropia geométrica pode ser resumida em um modelo de elipsóide em que haverá eixos de máximo, médio e mínimo alcance. A Figura 6.12 é uma representação da anisotropia geométrica.

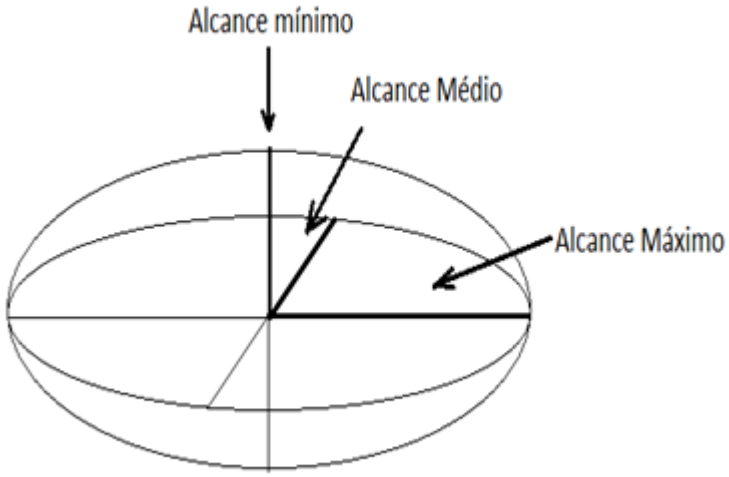


Figura 6.12: Representação do modelo de anisotropia geométrico. Elipsóide com valores de alcance mínimo médio e alcance máximo.

Os alcances em qualquer direção podem ser derivados de um modelo isotrópico unitário a partir de operações lineares, resultando em um novo sistema de coordenadas. A Equação 6.12 representa a matriz de rotação das coordenadas para os eixos de referência.

$$Q = \begin{bmatrix} cos\theta_3 & sin\theta_3 & 0 \\ -sin\theta_3 & cos\theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ cos\theta_2 & sin\theta_2 & 0 \\ -sin\theta_2 & cos\theta_2 & 0 \end{bmatrix} \begin{bmatrix} cos\theta_1 & sin\theta_1 & 0 \\ -sin\theta_1 & cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.12}$$

Em que θ_3 consiste no ângulo de rotação no eixo z, θ_2 o ângulo de rotação no eixo y e θ_1 a rotação do ângulo no eixo x. Os valores do vetor unitário são então redimensionados segundo a matriz de dilatação da Equação 6.13.

$$D = \begin{bmatrix} l_1 & 0 & 0 \\ 0 & l_2 & 0 \\ 0 & 0 & l_3 \end{bmatrix} \tag{6.13}$$

Em que l_1, l_2 e l_3 são os comprimentos dos eixos de máximo, médio e mínimo alcance. Tais matrizes são utilizadas para se construir o modelo do elipsoide de anisotropia e determinar os alcances em qualquer direção possível.

O caso zonal consiste em variações de patamares ao longo de direções diferentes. Demonstra-se o exemplo da anisotropia zonal. Observa-se na Figura 6.13 que a diferença de azimute pelo ângulo θ leva a uma diferença de patamares de g1 para g1 + g2. A anisotropia zonal é característica em alguns tipos de depósitos divididos em estratos, ao qual se verifica diferenças litológicas nas diversas camadas. A variabilidade na direção perpendicular aos estratos tende a ser diferente da direção paralela, que tende a ser mais contínua pelo princípio de sedimentação.

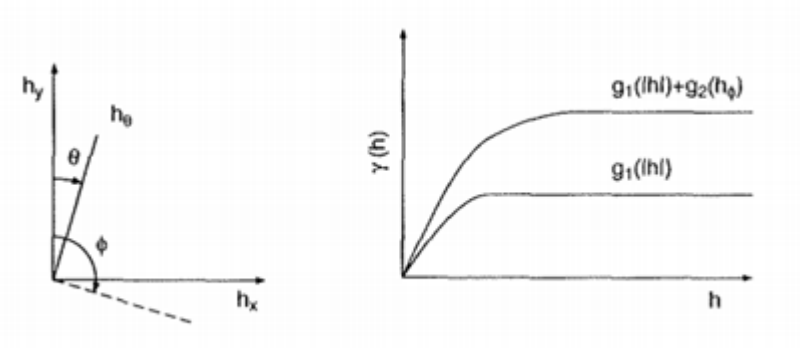


Figura 6.13: Anisotropia zonal representada nos variogramas a) Diferença entre as direções dos variogramas b) Representação da anisotropia por variogramas com patamares diferentes.

A modelagem de anisotropias zonais envolve a utilização de estruturas distintas de variograma que combinadas representarão o conjunto total. A anisotropia zonal também pode ser caracterizada como uma mudança de fenômeno em larga escala.

A anisotropia é uma flexibilização do modelo de variograma para atender às necessidades de depósitos mais complexos, que podem apresentar características diferenciadas segundo diversas direções.

6.5.5 Funções de continuidade espacial cruzadas

Na modelagem espacial de múltiplas variáveis, tal como os modelos de cokrigagem ou markovianos, é necessário determinar funções de continuidade cruzadas além das diretas. Estas não estão submetidas às mesmas condições de contorno das funções diretas. Primeiramente, porque o valor de patamar de uma estatística cruzada é sempre menor ou igual a das estatísticas diretas e está ligado à correlação entre as variáveis utilizadas para a modelagem.

O covariograma cruzado pode ser representada pela Equação 6.14 em que Zi e Zj são variáveis distintas e mi e mj são suas respectivas médias:

$$C_{ij}(h) = \frac{1}{2} E\left[\left(Z_i(x+h) - m_i\right)\left(Z_j(x) - m_j\right)\right] \tag{6.14}$$

O covariograma direto é uma função par e unicamente limitada por um valor de patamar. As funções cruzadas, no entanto, podem apresentar efeitos de retardo e não se comportarem como uma função par, tal que $C_{ij}(h) \neq C_{ij}(-h)$, e que i e j são variáveis aleatórias diferentes entre si. Toda função pode ser descrita como uma combinação de funções pares e ímpares. A covariância cruzada pode ser decomposta tal como na Equação 6.15:

$$C_{ij}(h) = \frac{1}{2}\left(C_{ij}(h) + C_{ij}(-h)\right) + \frac{1}{2}\left(C_{ij}(h) - C_{ij}(-h)\right) \tag{6.15}$$

Em que a soma de covariâncias representa o termo par da função e a diferença o termo ímpar. Há um sério problema em se definir a matriz de covariâncias, pois geralmente para um dado lag ela não poderá ser considerada nem positiva definida ou negativa definida.

Os efeitos produzidos pelo retardo não permitem a utilização de funções assimétricas na resolução dos sistemas de krigagem utilizados a posteriori. Segundo o mesmo autor, a dificuldade de caracterização da covariância no espaço de valores reais leva a utilização em números complexos.

O variograma cruzado, no entanto, não está sujeito aos efeitos do retardo tal como a covariância e apresenta unicamente um termo par. A Equação 6.16 expressa a fórmula da função:

$$\gamma_{ij}(h) = \frac{1}{2} E\left[\left(Z_i(x+h) - Z_i(x)\right)\left(Z_j(x+h) - Z_j(x)\right)\right] \tag{6.16}$$

6.5.6 Modelo linear de corregionalização

Segundo , o modelo linear de corregionalização implica que uma variável aleatória deve ser escrita como uma combinação linear de funções aleatórias independentes. Isso significa que para qualquer variável i e j, o modelo estrutural deve ser o mesmo, tal que $C(h) = \sum_{i=1}^{n} b_i \rho(h)$ e que $\rho(h)$ é um modelo único de correlograma e b_i é a contribuição para cada variável considerada. Para ser considerado um modelo permissível, o traço da matriz de covariância deve ser maior que a soma de qualquer coluna ou linha, ou que o determinante deva ser maior ou igual a zero. Os modelos lineares de corregionalização devem satisfazer a condição de matrizes positiva definidas para a resolução dos casos multivariados. A dificuldade de se estabelecerem modelos segundo os critérios necessários, levou à simplificações das krigagens colocadas e de modelos Markovianos.

6.5.7 Modelagem automática de variogramas

Na tentativa de minimizar o trabalho do avaliador na modelagem de funções cada vez mais complexas, a modelagem semiautomática também é uma alternativa para reduzir o erro do ajuste do modelo. Em 1985, já havia se iniciado a tentativa de modelagens automáticas por meio de mínimos quadrados ponderados. Em 1988, optou-se por utilizar alternativas não paramétricas no desenvolvimento de variogramas por transformadas de Fourier. A alternativa não paramétrica auxilia na obtenção rápida de mapas de variograma que representam a continuidade em um domínio espacial.

A necessidade de análises rápidas e eficientes aproximou a geoestatística cada vez mais da computação e dos algoritmos numéricos. O Varfit, um programa de uso livre para variogramas automáticos, constitui até hoje uma base de desenvolvimento para os softwares de modelagem automática em geoestatística. Houve modificações no programa para atender às necessidades do operador para pontos de âncora no variograma experimental. Estes pontos de âncora são valores do variograma experimental que possuem o ajuste coincidente com o seu valor naquele local.

Trabalhos mais atuais demonstram que a geoestatística preocupa cada vez mais em análises rápidas e menos laboriosas, tal como a utilização de variogramas automáticos juntamente com krigagem processada em múltiplos processadores em paralelo. Além disso, há a proposição de algoritmos interativos para a variografia.

O desenvolvimento dos recursos computacionais e de uma teoria mais abrangente permitiram o desenvolvimento de estudos em diversas áreas tal como na biologia, metalurgia entre outras áreas tais como também hidrogeologia, engenharia civil e ambiental.

O objetivo da modelagem automática de variogramas é criar um modelo consistente que envolva as principais características do fenômeno descrito, tais como anisotropia e comportamentos próximos da origem, sem a necessidade da interferência manual. A modelagem puramente computacional, sem interferência parcial do operador, leva à criação de continuidades artificiais pouco

representativas do fenômeno. A proposta semi-automática é então indicada, aos quais os eixos de maior, menor e média continuidade são definidos primordialmente.

Duas vertentes dos processos de otimização são descritas na bibliografia e se dividem em uma abordagem paramétrica e uma abordagem não paramétrica. Na primeira alternativa, propõem-se a otimização de funções já conhecidas e permissíveis, em contrapartida da segunda aos quais o ajuste é numérico e não é estipulada uma função propriamente dita.

As propostas desenvolvidas a partir da década de setenta constam desde a metodologia de mínimos quadrados, pela utilização de valores ponderados, ou por métodos que envolvam hipótese de multi-gaussianidade. Na sua grande maioria, os métodos de modelagem automática são definidos pelo modelo que levar ao menor desvio médio quadrático.

Há a necessidade da utilização de ponderadores para os diversos pontos experimentais para o ajuste de variogramas, à medida que para distâncias mais curtas é necessário um melhor ajuste . As alternativas propostas indicam a utilização do número de pares da estatística, o inverso da distância e do valor do variograma experimental como medidas de ajuste. Mesmo definindo pesos para os valores experimentais, a modelagem automática ainda pode requerer intervenção do operador.

Em todos os modelos de otimização do ajuste de variogramas, é necessário construir uma função objetivo que é responsável pela aproximação dos valores estimados e dos experimentais. Geralmente, procura-se otimizar a dissimilaridade entre os valores conjugados. A Equação 6.17 demonstra a relação de dissimilaridade entre o modelo e os variogramas experimentais:

$$\psi = \sum_{i=0}^{n} \rho_i(\gamma_i - \gamma_i^*) \tag{6.17}$$

Em que ψ é a equação objetivo, γ_i são os valores experimentais e γ_i^* são os valores de um modelo a ser ajustado, para uma função de ajuste ρ.

Exercise 6.1 A tabela a seguir determina amostras segundo uma direção X e seus valores de teor associados. Determine:

1. O variograma experimental para um lag igual a 1
2. O variograma experimental para um lag igual a 2
3. A covariância experimental para um lag igual a 1
4. A covariância experimental para um lag igual a 2

x	teores
1	0.8
2	0.75
3	0.7
4	0.75
5	0.8
6	0.85
7	0.9
8	0.85
9	0.8
10	0.75
11	0.7

7. Krigagem

7.1 Introdução

A krigagem é um termo genérico que expressa um conjunto de metodologias de estimativa que levam em consideração o mínimo valor do erro. Os métodos também são chamados de BLUE (Best linear umbiesed estimation). Entre os procedimentos podemos citar a krigagem ordinária, krigagem simples, krigagem da probabilidade e krigagem universal. Algumas metodologias não lineares tais como a krigagem de indicadores e a krigagem gaussiana também recebem o mesmo nome, pois utilizam operações lineares em dados transformados.

É importante entender, antes de tudo, que estimar é um processo sempre associado ao erro. O que é possível de se fazer durante uma estimativa é sempre reduzí-lo ao máximo e encontrar o valor mais provável de ocorrência. No entanto, se um evento tem baixa probabilidade, como ganhar em uma sena, ainda sim há pessoas que por hora ganham no jogo. Da mesma forma se algo possui grande probabilidade de ocorrer, como um lutador de boxe ganhar em uma luta contra um menino de cinco anos, ainda sim podemos nos surpreender.

Na krigagem utilizamos um estimador para determinar a realização de uma variável aleatória em um ponto desconhecido. Este é uma combinação linear de vários valores de amostras ao redor do ponto a ser estimado (7.1)

$$z_0 = \sum_{i=1}^{n} \lambda_i z_i \tag{7.1}$$

Em que λ_i são valores de peso associados a cada uma das amostras utilizadas para a estimativa. Não utilizamos todas as amostras do domínio, porque primeiramente, causará uma grande suavização nas estimativas e em segundo, porque computacionalmente é melhor realizar o filtro nos dados e inverter matrizes de krigagem pequenas, do que inverter matrizes de krigagem grandes. O tempo de krigagem de subproblemas não é diretamente proporcional a de grandes problemas.

Como demonstrado no capíulo um podemos definir o erro de estimativa segundo a equação

(7.2)

$$\varepsilon(z_0^*) = z_0^* - z_0 \tag{7.2}$$

Em que z_0^* é o valor estimado no ponto desconhecido e z_0 é o valor real naquele ponto. Substituindo a equação (7.1) em (7.2) obtemos a relação (7.3)

$$\varepsilon(z_0^*) = \sum_{i=1}^{n} \lambda_i z_i - z_0 \tag{7.3}$$

Tomando o quadrado do erro de estimativa temos a seguinte a relação (7.4)

$$\varepsilon(z_0^*)^2 = \sum_{i=1}^{n} \lambda_i \sum_{i=j}^{n} \lambda_j z_i z_j - 2\sum_{i=1}^{n} \lambda_i z_i z_0 - z_0^2 \tag{7.4}$$

Tomando o menor valor esperado do erro quadrádico temos então

$$E\left(\varepsilon(z_0^*)^2\right) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \lambda_j Cov(z_i, z_j) - 2\sum_{i=1}^{n} \lambda_i Cov(z_i, z_0) - Cov(z_0, z_0) \tag{7.5}$$

Que também é chamada variância de extensão. Ou seja, esta é uma estimativa de quanto varia tomarmos como valor de um ponto desconhecido uma combinação linear de valores mais próximos dele. Para encontrarmos a variância de krigagem precisamos minimizar esta variância de extensão, adicionando a condição de não viés da estatística. Como demonstrado no capítulo 5 no subitem 5.1 a condição de não viés amostral neste caso é que a soma dos ponderadores deve ser igual a 1.

Adicionando a restrição no problema e tomando as derivadas parciais para cada um das equações consideradas temos a relação demonstrada por (7.5)

$$\sigma_{krig}^2 = \frac{\partial}{\partial \lambda_i}\left(\sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \lambda_j Cov(z_i, z_j) - 2\sum_{i=1}^{n} \lambda_i Cov(z_i, z_0) - Cov(z_0, z_0)\right) + \frac{\partial}{\partial \lambda_i} R = 0 | \forall i \tag{7.6}$$

Em que R é a restrição de não enviesamento adicionado ao problema, associada a soma dos ponderadores da krigagem. Tomando o valor das derivadas parciais temos o seguinte conjunto de equações (7.7)

$$\sigma_{krig}^2 = 2\sum_{i=1}^{n} \lambda_i Cov(z_i, z_j) - 2Cov(z_i, z_0) + \frac{\partial}{\partial \lambda_i} R = 0 | \forall i \tag{7.7}$$

O cálculo das derivadas parciais pode ser realizado de acordo com a expansão dos somatórios como demonstrado na equação (7.5)

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_j \lambda_i Cov(z_i, z_j) = \sum_{j=2}^{n} \lambda_j \lambda_1 Cov(z_1, z_j) + \sum_{i=2}^{n} \lambda_1 \lambda_i Cov(z_i, z_1) + \lambda_1 \lambda_1 Cov(z_1, z_1) \tag{7.8}$$

Cada sistema de krigagem pode então ser resolvido de acordo com uma matriz de covariâncias genérica como descrito em (7.6)

$$\begin{pmatrix} Cov(z_1,z_1) & Cov(z_1,z_2) & ... & Cov(z_1,z_n) & \frac{\partial R}{\partial \lambda_1} \\ Cov(z_2,z_1) & Cov(z_2,z_2) & ... & Cov(z_2,z_n) & \frac{\partial R}{\partial \lambda_2} \\ ... & ... & ... & ... & ... \\ Cov(z_n,z_1) & Cov(z_n,z_2) & ... & Cov(z_n,z_n) & \frac{\partial R}{\partial \lambda_n} \\ 1 & 1 & 1 & ... & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \\ 1/2 \end{pmatrix} = \begin{pmatrix} Cov(z_0,z1) \\ Cov(z_0,z2) \\ ... \\ Cov(z_0,zn) \\ P \end{pmatrix} \tag{7.9}$$

Em que P é o valor da restrição para a soma dos ponderadores. O termo da esquerda da matriz é responsável pelo desagrupamento dos dados, enquanto o termo da direita é responsável por ponderar a distância do ponto estimado até a amostra considerada. Esse sistema de matrizes proposto aqui é genérico, e qualquer krigagem pode ser descrita a partir dele, bastando apenas considerar diferentes variáveis e funções de restrição ao enviesamento R e o valor P de restrição à soma dos ponderadores. Nos tópicos a seguir demonstraremos as restrições quanto a krigagem ordinária e simples. Encontrados os pesos da krigagem podemos encontrar o valor estimado pela equação (7.2)

7.2 Krigagem Ordinária

Para a krigagem ordinária utilizamos os mesmos pressupostos e cálculos utilizados em 7.1. Logo temos como restrição à condição de não viés dado pela demonstração abaixo:

Demonstração. $Z_0 = \sum_{i=1}^{n} \lambda_i Z_i$
$E(Z_0) = E\left(\sum_{i=1}^{n} \lambda_i Z_i\right) = m$
$E(Z_0) = \sum_{i=1}^{n} E\left(\lambda_i Z_i\right) = m$
$E(Z_0) = \sum_{i=1}^{n} \lambda_i E\left(Z_i\right) = m$
$E(Z_0) = \sum_{i=1}^{n} \lambda_i m = m$
$m \sum_{i=1}^{n} \lambda_i = m$
$\sum_{i=1}^{n} \lambda_i = 1$

■

Logo nossa função de restrição pode ser determinada por (7.10), e o nosso valor P é igual a 1.

$$R_i = \mu \left(\sum_{i=0}^{n} \lambda_i - 1 \right) | \forall i \tag{7.10}$$

Em que μ é o multipliador langragiano. Logo tomando a derivada parcial de cada uma das restrições para cada uma das amostras i do problema temos a relação segundo a equação (7.11)

$$\frac{\partial}{\partial \lambda_i} R_i = \mu | \forall i \tag{7.11}$$

O sistema de equações da krigagem pode ser transformado então na relação 7.12

$$\begin{pmatrix} Cov(Z_1,Z_1) & Cov(Z_1,Z_2) & ... & Cov(Z_1,Z_n) & 1 \\ Cov(Z_2,Z_1) & Cov(Z_2,Z_2) & ... & Cov(Z_2,Z_n) & 1 \\ ... & ... & ... & ... & ... \\ Cov(Z_n,Z_1) & Cov(Z_n,Z_2) & ... & Cov(Z_n,Z_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} Cov(Z_0,Z_1) \\ Cov(Z_0,Z_2) \\ ... \\ Cov(Z_0,Z_n) \\ 1 \end{pmatrix} \tag{7.12}$$

7.3 Krigagem Simples

A krigagem simples tem como pressuposto encontrar os ponderadores que minimizem o resíduo da variável aleatória. Como determinado em 2.4, nada mais é que a própria variável subtraído do valor médio da função aleatória. Logo o método requer antes de tudo conhecimento do valor médio e da hipótese de estacionaridade de segunda ordem. Neste caso temos que o valor estimado pode ser descrito pela equação (7.13)

$$Z_0^* = \sum_{i=0}^{n} \lambda_i (Z_i - m) + m \tag{7.13}$$

Podemos isolar os termos da equação (7.7) em relação ao valor médio assim obtendo a equação (7.14)

$$Z_0^* = \sum_{i=0}^{n} \lambda_i Z_i + m \left(1 - \sum_{i=0}^{n} \lambda_i\right) \tag{7.14}$$

Ou seja, notamos que na krigagem simples parte dos pesos é atribuído à variável aleatória e parte para a média global. Considerando a condição de não viés amostral temos que a soma dos ponderadores deve ser igual a zero

Demonstração. $E(Z_0^*) = E(\sum_{i=0}^{n} \lambda_i Y_0 + m) = m$
$E(Z_0^*) = \sum_{i=0}^{n} \lambda_i E(Y_0) + E(m) = m$
$\sum_{i=0}^{n} \lambda_i E(Y_0) + m = m$
$\sum_{i=0}^{n} \lambda_i E(Y_0) = 0$
$E(Y_0) = 0 \vee \sum_{i=0}^{n} \lambda_i = 0$ ■

Caso a escolha da média da função aleatória seja realmente correta e o caso perfeitamente estacionário nenhuma condição seria necessária para o não enviesamento da estatística. No entanto, para forçarmos o sistema de resolução das matrizes de krigagem encontrar valores condizentes optamos por adicionar a condição de que a soma dos ponderadores deve ser igual a zero. Em outras palavras, diferentemente da krigagem ordinária, a krigagem simples pressupõe o conhecimento intrínseco da média da função aleatória, o que na maioria das vezes não é realidade.

Nossa função de restrição se torna portanto a equação (7.15) e o nosso valor P de restrição à soma dos ponderadores é igual a 0.

$$R_i = \mu \sum_{i=0}^{n} \lambda_i | \forall i \tag{7.15}$$

Em que μ é o multiplicador lagrangiano. Tomando a derivada parcial de cada restrição para cada índice temos que

$$\frac{\partial}{\partial \lambda_i} R_i = \mu | \forall i \tag{7.16}$$

Logo o sistema de matrizes para a krigagem simples pode ser transformado em (7.17)

$$\begin{pmatrix} Cov(Y_1,Y_1) & Cov(Y_1,Y_2) & \dots & Cov(Y_1,Y_n) & 1 \\ Cov(Y_2,Y_1) & Cov(Y_2,Y_2) & \dots & Cov(Y_2,Y_n) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(Y_n,Y_1) & Cov(Y_n,Y_2) & \dots & Cov(Y_n,Y_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} Cov(Y_0,Y_1) \\ Cov(Y_0,Y_2) \\ \dots \\ Cov(Y_0,Y_n) \\ 0 \end{pmatrix} \tag{7.17}$$

Sendo a covariância dos resíduos a mesma covariância das amostras. Calculado os pesos de cada um dos resíduos encontramos o valor estimado.

7.4 Krigagem de blocos

Um caso especial de krigagem ocorre quando o variável estimada tem um suporte diferente das amostras. A forma mais simples de se resolver este problema é estimando uma série de valores dentro do suporte a ser estimado e tomando seu valor médio. A figura (7.1) demonstra um bloco B contendo novo pontos k1 até k9 contidos nele.

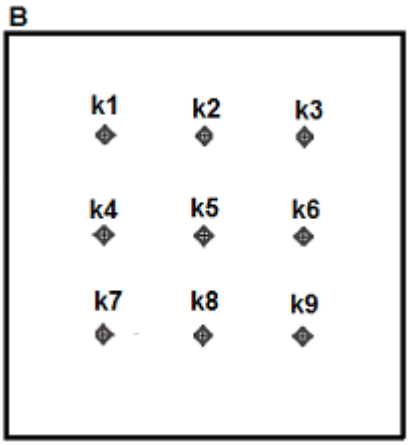


Figura 7.1: Demonstração dos pesos de krigagem para o posicionamento de amostras para um modelo de pepita puro

Neste caso o valor da variável aleatória para o suporte estimado pode ser dado por uma média das combinações lineares de variáveis pontuais contidas dentro da região a ser estimada tal como em (7.18)

$$Z_v = \frac{1}{p}\sum_{j=1}^{p} Z_j \tag{7.18}$$

Em que p é o número de variáveis contidas dentro daquele volume. Logo temos o erro de estimativa dado pela equação (7.19)

$$\varepsilon(Z_v^*) = Z_v^* - Z_v \tag{7.19}$$

Demonstração. $\varepsilon(Z_v^*) = \frac{1}{p}\sum_{j=1}^{p}(\sum_{i=1}^{n}\lambda_i Z_i - Z_j)$
$\varepsilon(Z_v^*) = \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{n}\lambda_i Z_i - \frac{1}{p}\sum_{j=1}^{p} Z_j$
$\varepsilon(Z_v^*) = \frac{1}{p}p\sum_{i=1}^{n}\lambda_i Z_i - \frac{1}{p}\sum_{j=1}^{p} Z_j$
$\varepsilon(Z_v^*) = \sum_{i=1}^{n}\lambda_i Z_i - \frac{1}{p}\sum_{j=1}^{p} Z_j$
$\varepsilon(Z_v^*)^2 = \sum_{i=1}^{n}\sum_{i'=1}^{n}\lambda_i\lambda_{i'}Z_iZ_{i'} - \frac{2}{p}\sum_{i=1}^{n}\sum_{j=1}^{p}\lambda_i Z_iZ_j + \frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'=1}^{p} Z_jZ_{j'}$
$E\left(\varepsilon(Z_v^*)^2\right) = \sum_{i=1}^{n}\sum_{i'=1}^{n}\lambda_i\lambda_{i'}Cov(Z_i,Z_{i'}) - \frac{2}{p}\sum_{i=1}^{n}\sum_{j=1}^{p}\lambda_i Cov(Z_iZ_j) + \frac{1}{p^2}\sum_{j=1}^{p}\sum_{j'=1}^{p} Cov(Z_jZ_{j'})$
Derivando em relação ao ponderador tal como demonstrado na seção anterior encontramos a seguinte equação para a variância de krigagem:
$\sigma^2_{krig} = 2\sum_{i'=1}^{n}\lambda_{i'}Cov(Z_i,Z_{i'}) - \frac{2}{p}\sum_{j=1}^{p} Cov(Z_iZ_j)|\forall i$

Em que:
$\overline{Cov}(Z_iZ_j) = \frac{1}{p}\sum_{j=1}^{p} Cov(Z_iZ_j)$
Logo temos:
$\sigma^2_{krig} = 2\sum_{i=1'}^{n} \lambda_i Cov(Z_i, Z_{i'}) - 2\overline{Cov}(Z_iZ_j) | \forall i$

■

Ou seja, se estamos estimando um bloco a partir de um ponto, a única diferença entre o sistema de krigagem convencional é que o lado da direita da matriz é substituído pela média das covariâncias entre cada amostra e os pontos estimados dentro do bloco. A figura (7.2) demonstra como deve-se proceder para calcular a covariância média para cada amostra na krigagem de blocos.

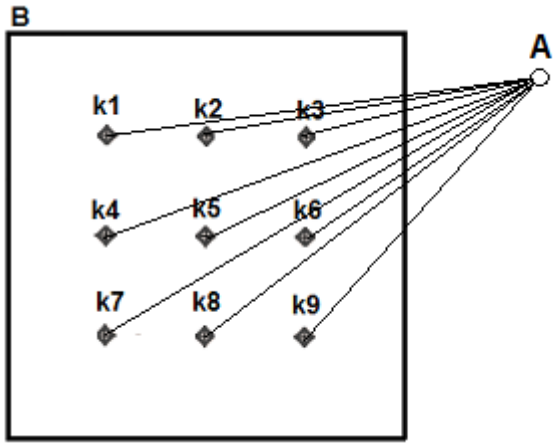


Figura 7.2: Covariância média como a média de covariâncias entre cada ponto estimado k dentro do bloco e o valor de cada amostra A

Logo o sistema de krigagem para blocos é demonstrado em (7.20)

$$\begin{pmatrix} Cov(Z_1,Z_1) & Cov(Z_1,Z_2) & \dots & Cov(Z_1,Z_n) & 1 \\ Cov(Z_2,Z_1) & Cov(Z_2,Z_2) & \dots & Cov(Z_2,Z_n) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(Z_n,Z_1) & Cov(Z_n,Z_2) & \dots & Cov(Z_n,Z_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} \overline{Cov}(Z_0,Z_1) \\ \overline{Cov}(Z_0,Z_2) \\ \dots \\ \overline{Cov}(Z_0,Z_n) \\ 1 \end{pmatrix} \quad (7.20)$$

7.5 Influência nos pesos da krigagem

A krigagem é na verdade um estimador que não leva em consideração o valor da ammostra, mas apenas sua correlação espacial e disposição no espaço das amotras. Essa é a grande crítica aos métodos de krigagem que atualmente estão sendo substituídos aos poucos pelos métodos de simulação geoestatística. Mostraremos a influência nos pesos da krigagem quanto a disposição espacial quanto ao modelo de continuidade espacial adotado e quanto ao posicionamento das amostras.

7.5.1 Influência do modelo de continuidade espacial nos pesos

A figura (7.3) demonstra o posicionamento de quatro amostras em relação ao ponto estimado. As amostras no sentido vertical da figura estão mais próximas que no sentido horizontal. Quando

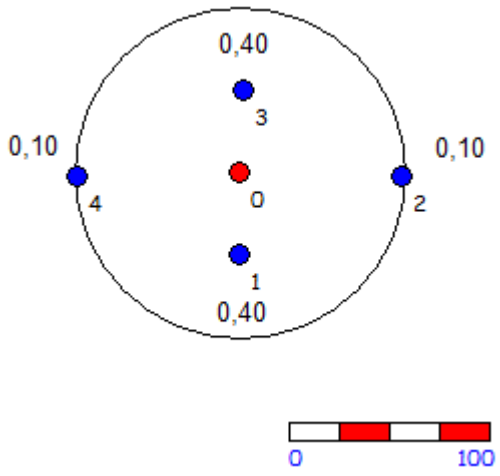


Figura 7.4: Demonstração dos pesos de krigagem para o posicionamento de amostras em um modelo esférico com alcance igual a 2L. A linha cheia representa o alcance do variograma

considerado um modelo de pepita puro, os pesos de krigagem são sempre os mesmos independente do posicionamento das amostras.

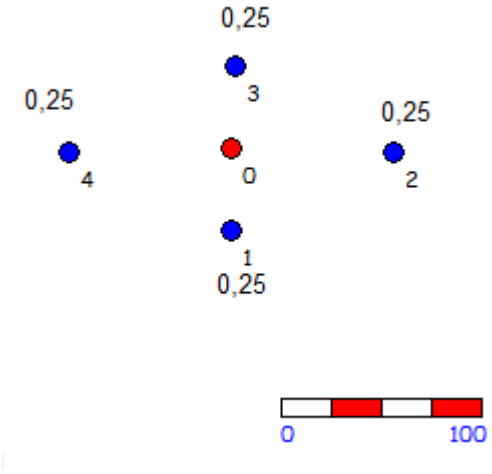


Figura 7.3: Demonstração dos pesos de krigagem para o posicionamento de amostras para um modelo de pepita puro

A figura (7.4) demonstra o mesmo caso para um modelo esférico. O range adotado é igual a 2L. O modelo esférico e exponencial tende a dar pesos diferenciados para as amostras mais próximas, tal que seu valor é maior quanto mais próxima for a amostra do ponto a ser estimado.

A figura (7.5) demonstra o mesmo caso para um modelo gaussiano. O range prático adotado é de 1.5 L. O modelo gaussiano é sem dúvida o mais suavizador de todos os outros modelos para a krigagem, considerando seu comportamento parabólico próximo à origem. Maiores pesos são atribuídos às distâncias mais proximais.

7.5.2 Influência dos parâmetros do variograma

O alcance do variograma altera os pesos de krigagem aumentando os valores das amostras mais próximas para um aumento do mesmo. A figura (7.6) demonstra os pesos de krigagem para um

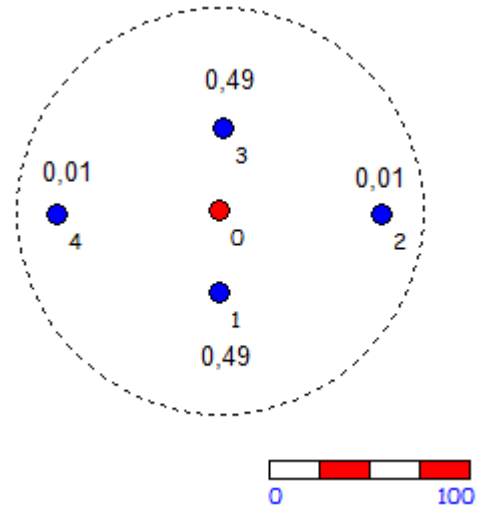


Figura 7.5: Demonstração dos pesos de krigagem para o posicionamento de amostras em um modelo gaussiano com alcance igual a 1.5L. A linha hachurada demonstra o alcance prático do modelo de continuidade espacial.

modelo esférico de variograma com um alcance de 63m e outro de 125m. As amostras estão dispostas em uma cruz com o ponto estimado tal que o eixo maior possui 162m e o eixo menor igual a 80m.

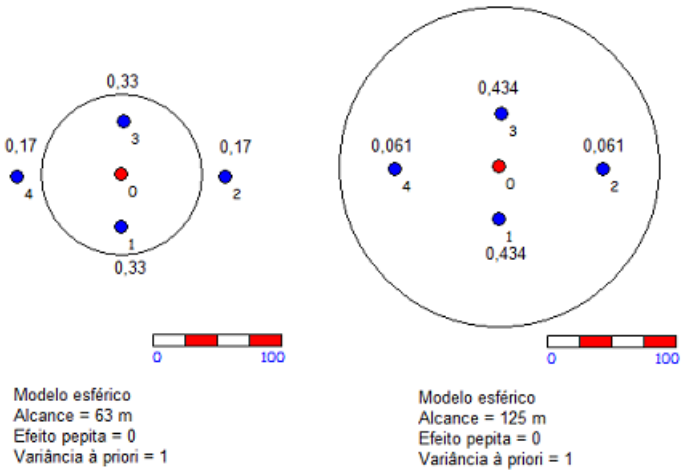


Figura 7.6: Influência dos parâmetros do variograma na krigagem. Efeito do alcance. Maiores alcances atribuem mais peso à amostras mais próximas

Quanto a variância à priori do variograma notamos segundo a figura (7.7) que qualquer valor adicionado não altera os pesos de krigagem. No entanto, apesar de não alterar os pesos, um aumento na variância à priori causa um aumento na variância de krigagem.

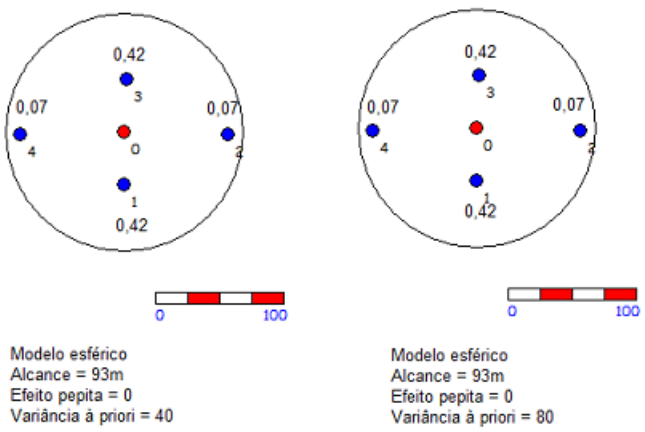


Figura 7.7: Influência dos parâmetros do variograma na krigagem. Efeito da variância a priori. Maiores valores de variância não alteram os pesos das amostras

Em último caso notamos segundo a figura (7.8) que o efeito pepita tende a normalizar os pesos de krigagem. Quanto maior for o efeito pepita e menor a contribuição, mais o modelo de variograma se aproxima de efeito pepita puro e neste caso qualquer geometria das amostras produzirá pesos iguais.

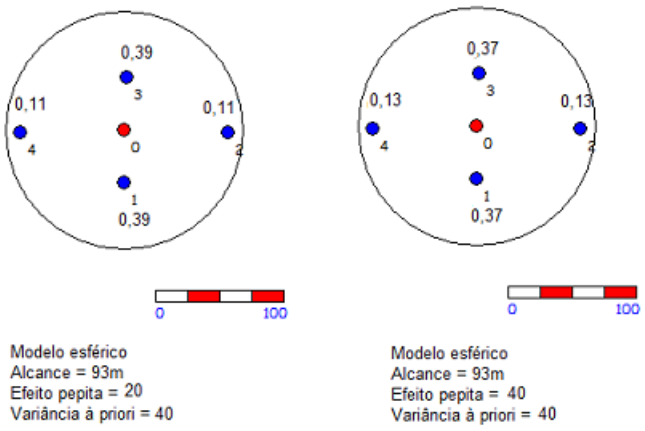


Figura 7.8: Influência dos parâmetros do variograma na krigagem. Efeito do alcance. Maiores alcances atribuem mais peso à amostras mais próximas

7.5.3 Efeito da geometria das amostras

Quanto a geometria das amostras é necessário lembrar que distâncias iguais entre as amostras e o ponto estimado produzem pesos iguais, no caso de um modelo de continuidade espacial isotrópico. A figura (7.9) demonstra esta relação.

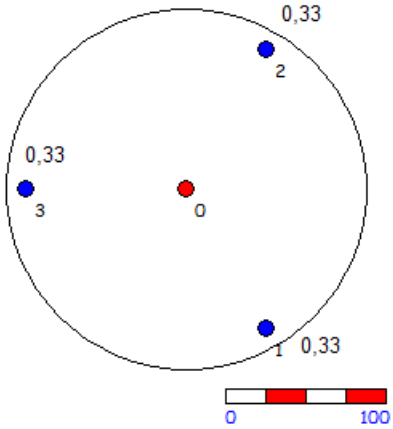


Figura 7.9: Influência no peso para distâncias iguais das amostras ao ponto estimado. Pesos iguais para um modelo de continuidade espacial isotrópico.

Para amostras agrupadas a tendência é produzir pesos iguais. A figura (7.10) demonstra esta situação.

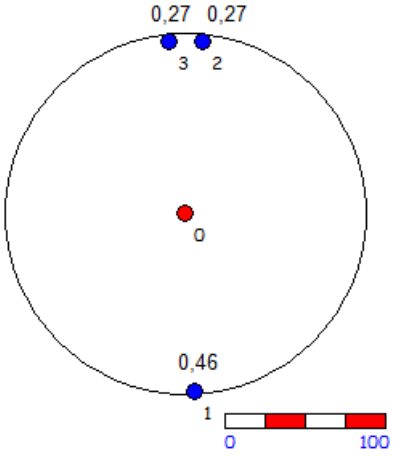


Figura 7.10: Influência de amostras agrupadas nos pesos da krigagem. Tendência de valores iguais para os pesos.

O caso mais importante para o posicionamento das amostras é quando existem amostras à frente de outras. Nesse caso ocorre "blindagem"e é possível que elas recebam pesos negativos. O termo em inglês para isto é "screen effect". Pesos negativos para um valor estimado são comuns de ocorrer, o que não é adequado muitas vezes é um valor negativo para estimativas, quando a variável somente pode assumir valores positivos. Neste caso é importante um controle sobre a estratégia de busca de forma a garantir a melhor situação para ponderadores negativos. A figura (7.11) demonstra o feito de blindagem das amostras e associação com pesos negativos.

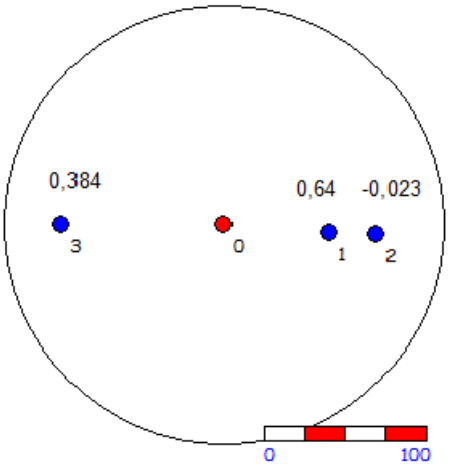


Figura 7.11: Influência da blindagem de amostras. Pesos negativos para amostras que estão encobertas por outras.

Outra variável importante que influencia no peso da krigagem é o suporte do valor estimado. Neste caso estamos lidando com um tipo diferente de krigagem também chamada de krigagem de blocos ao qual o suporte estimado é diferente do suporte das amostras. Blocos maiores tendem neste caso a normalizar o peso das amostras, mas nunca a igualá-los tal como acontece com o modelo de efeito pepita puro. A figura (7.12) tende a demonstrar esta situação.

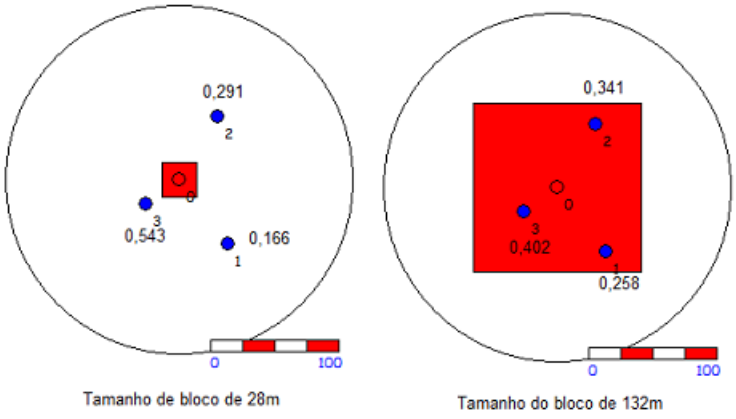


Figura 7.12: Influência do efeito de suporte do valor estimado. Blocos maiores tendem a normalizar o peso das amostras.

7.6 Estratégia de procura

Como dito anteriormente a krigagem exige que determinemos uma região envolta do ponto a ser estimado que determinará as amostras que ponderarão a estimativa. Escolher uma região muito pequena poderá fazer com que o algoritmo não encontre nenhuma amostra. Escolher uma região muito maior poderá suavizar a estimativa a ponto de tornar o valor da amostra muito próximo da média global. A escolha da estratégia de busca ideal é um fator muito mais influente muitas vezes do que um ajuste mais apurado no modelo de continuidade espacial.

Diversas geometrias podem ser escolhidas para se realizar a estratégia de procura dos pontos, mas as mais importantes são sem dúvida a circular e a elíptica. Buscas em caixa também podem ser feitas.

Dentre os parâmetros de krigagem mais utilizados são:

- Número máximo e número mínimo de amostras dentro da região de busca
- Forma, dimensão e orientação da região de busca
- Uso de estratégia de busca por octantes

Na maioria dos algoritmos de krigagem, a escolha definido a região de busca, caso o ponto a ser estimado esteja envolta de um número menor que o mínimo de amostras ou máximo aquele ponto simplesmente não é estimado.

Quanto a forma, dimensão e orientação da região de busca é importante lembrar que dimensões muito grandes produzirão grande suavização nos valores krigados, o que pode não corresponder com a realidade. Valores muito pequenos, no entanto, podem atribuir poucos pontos na estimativa e torná-la também inadequada.

Uma das formas utilizadas para reduzir a suavização da krigagem é utilizar uma estratégia de busca elíptica perpendicular à continuidade espacial dos dados. Dessa forma temos duas forças contrárias agindo para garantir maior homogeneidade aos pesos e maior erraticidade aos valores estimados. A figura (7.13) demonstra esta situação.

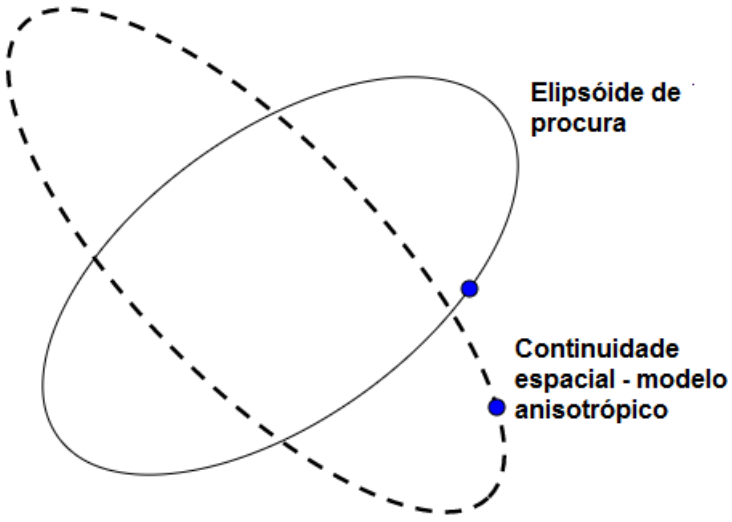


Figura 7.13: Estratégia de busca perpendicular à continuidade do fenômeno. Forma utilizada para garantir maior erraticidade aos dados estimados.

Outra forma de se realizar a estratégia de busca para a krigagem é utilizando octantes. Dessa forma podemos estipular um número mínimo ou máximo de amostras por cada octante para ser utilizado durante a krigagem. Uma conduta coerente é ser condizente com o número mínimo e máximo de amostras utilizada para a krigagem em cada octante. Se em uma estimativa usa-se um número mínimo de 8 amostras para a krigagem e 32 como o máximo, é plausível escolher uma estratégia de octantes que utilize um mínimo de 1 amostra por octante e no máximo 4. A figura (7.14) demonstra a estratégia de octantes para o caso considerado. Se o número mínimo de amostras por octante fosse de 3 amostras, os octantes 1,4,6 e 7 estariam descartados.

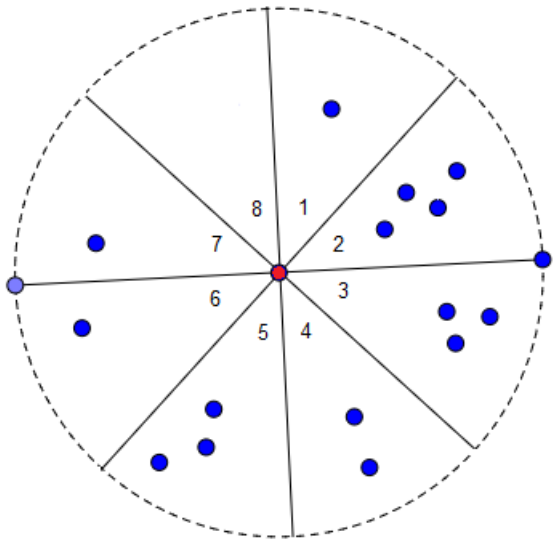


Figura 7.14: Estratégia de busca por octantes. Octantes 1,4,6 e 7 descartados por não possuírem o mínimo de amostras igual a 3

7.7 Validação da krigagem

Após realizada a krigagem devemos investigar se os valores estimados estão realmente próximos do esperado. Algumas metodologias são utilizadas para isso, entre elas citamos:

1. Verificação do comportamento dos mapas krigado e das amostras
2. Comparação da média global com a média das amostras
3. Análise de deriva de bandas do mapa
4. Validação Cruzada
5. Verificação de pesos negativos

7.7.1 Verificação do comportamento dos mapas krigado e das amostras

É importante que o mapa de valores krigados apresente comportamento semelhante das amostras. Nesta etapa verifica-se se a continuidade dos dados estimados é visualmente condizente com as características do fenônmeno estudado, tal como continuidade espacial e regiões de maior ou menor valor da variável aleatória.

A figura (7.15) demonstra a comparação visual entre um mapa realizado por polígonos de influência da amostra e outro krigado. Nota-se que em ambos a continuidade espacial dos dados apresenta direção NW e que as regiões de maior e menor valor são semelhantes.

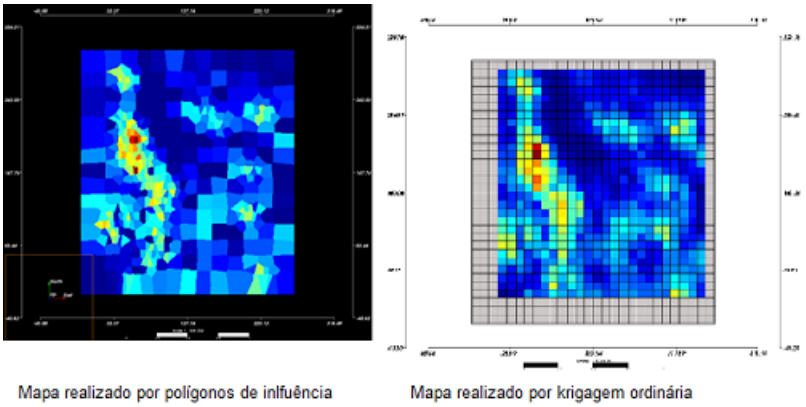


Figura 7.15: Comparação visual entre os valores da amostra por polígono de influência e dos blocos krigados.

7.7.2 Comparação da média global com a média das amotras

A krigagem é uma estatística não enviesada, isso significa que a média das amostras deve ser idêntica à media dos valores krigados. No entanto, a variância dos valores krigados é menor que a variância das amostras devido o efeito de suporte.

7.7.3 Análise de deriva de bandas do mapa

Não obstante a média das amostras deve ser a média do valor krigado, a média em subdomínios do mapa krigado deve ser igual a média dos subdomínios das amostras. Para isso realizamos um gráfico como demonstrado na figura (7.16). Dividimos o domínio espacial das amostras e dos valores krigado em bandas e tomamos os valores médios de cada banda colocando em um gráfico. Se o comportamento das duas curvas for semelhante falhamos em aceitar a hipótese de deriva nos dados. As bandas no gráfico podem ser analisadas nas diferentes direção de independência dos eixos cartesianos.

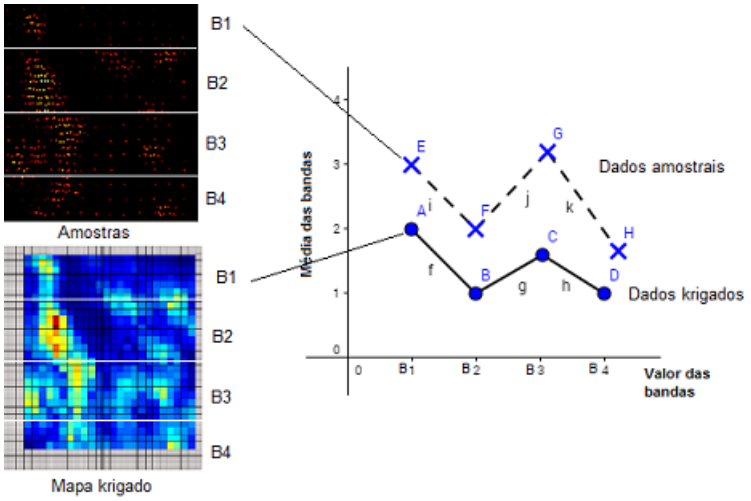


Figura 7.16: Exemplo da análise de deriva. Mama dos valores krigados e das amostras dividido em bandas e média tomaada para cadab banda em cada mapa demonstrado ao lado. Comportamento do gráfico semelhantne para as duas situações. Descarte da hipótese de deriva.

7.7.4 Validação cruzada

A validação cruzada é uma estimativa do erro de krigagem possível. Para isso retiramos um ponto amostral e estimamos sem aquele dado no ponto novamente. A diferença entre o valor estimado e o valor real da amostra é uma medida de erro, com média zero e variância determinada pelo conjunto de amostras. A validação cruzada não é necessariamente um valor de erro real cometido pelo método, mas uma alternativa comparativa entre diferentes modelos de continuidade espacial e estratégia de busca. Antes de realizar a krigagem, é interessante testar valores de validação cruzadas diferentes para encontrar a melhor estratégia de busca.

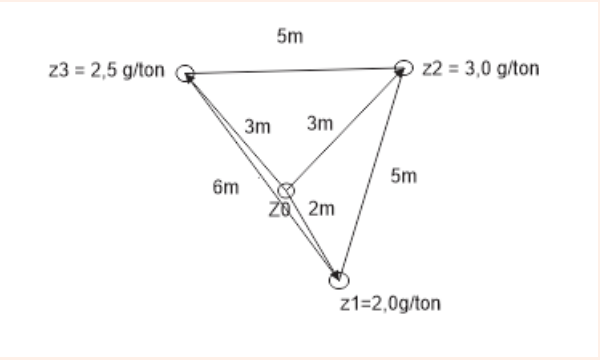
7.7.5 Verificação de pesos negativos

Como dito nas seções anteriores amostras blindadas podem produzir pesos negativos na krigagem. É interessante controlar a quantidade de pesos negativo na estimativa a fim de produzir um resultado com menor suavização. Para reduzir a quantidade de pesos negativos opta-se por reduzir o tamanho da região de busca de amostras na krigagem.

Exercise 7.1 A figura abaixo demonstra 3 pontos amostras e um ponto a ser estimado z0. Considere o modelo de continuidade espacial como:

$$gamma(h) = \begin{pmatrix} 12\left(\left[\frac{3h}{4}\right] - \left[\frac{h^3}{64}\right]\right), h \leqslant 4 \\ 12, h > 4 \end{pmatrix}$$

Determine os pesos de krigagem para cada uma das amostras e determine o valor estimado no ponto Z0.



8. Mudança de suporte

8.1 Mudança de suporte

Após realizada a krigagem dos dados é interessante comparar o efeito da suavização da krigagem. Esta possui duas forças envolvendo a estimativa em um local, uma relacionada com a interpolação dos dados e outra com a suavização dada pela tomada de valores médios. A fim de comparar as estatísticas estimadas com as amostras podemos tranformar a primeira em um suporte pontual. Logo a diferença entre os seus valores será apenas o efeito da suavização. Como demonstrado no capítulo 1 na figura (8.1), a variância do valor médio tende a estreitar com o aumento do suporte utilizado. Para comparar os histogramas precisamos então realizar um "esticamento"da distribuição.

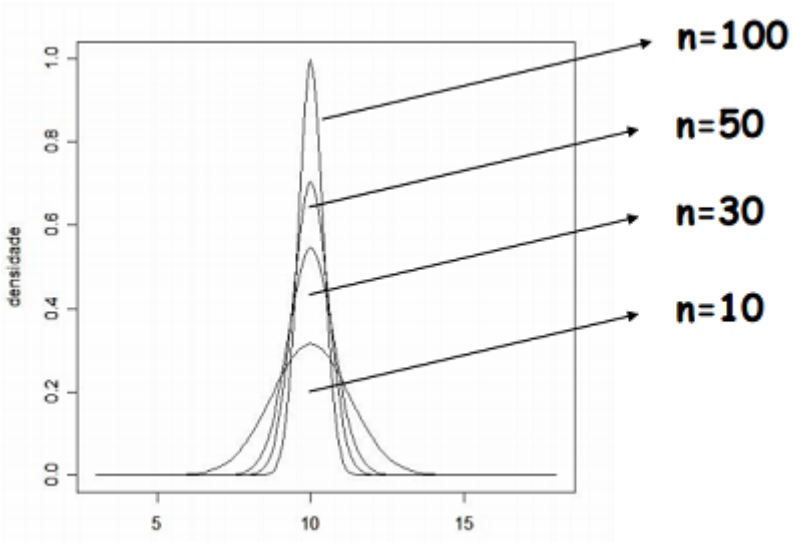


Figura 8.1: Figura demonstrando o efeito de suporte para um número crescente de amostras. O aumento do número de amostras tende a concentrar a função de densidade de probabilidade entorno do valor médio

Duas premissas devem ser tomadas antes de se realizar a correção de suporte. A primeira é de que o valor médio permanece constante. A segunda é que a variância da distribuição é corrigida por um fator "f". Esse fator f pode ser descrito pela equação (8.1)

$$f = \frac{\sigma_0^2 - \bar{\gamma}(V,V)}{\sigma_0^2} \tag{8.1}$$

Em que $\bar{\gamma}(V,V)$ é o valor de variograma médio dentro do suporte V a ser corrigido e σ_0^2 é o valor de variância à priori dos dados.

8.1.1 Correção afim

Uma das formas mais simples de se realizar a correção de suporte é mudando a distribuição de probabilidades por um fator linear. Essa também é chamada de "affine correction"ou correção afim, em que o valor da média da distribuição é mantida constante mas a variância sofre extensões de igual valor dado pela equação (8.2)

$$q' = \sqrt{f} * (q - m) + m \tag{8.2}$$

em que q é o quartil a ser transformado, m o valor médio e f o fator de correção demonstrado na seção anterior. Essa transformação linear proposta pelo método produz em alguns casos valores anômalos principalmente nas terminações das distribuições que apresentam comportamento muito mais assintótico que os valores centrais.

8.1.2 Transformação lognormal indireta

De forma a corrigir o comportamento das distribuições de forma mais assintótica nas terminações da distribuição, o método de transformação lognormal indireta propõe uma resoulção não linear para o problema, tal que cada quartil pode ser dado pela equação (8.3)

$$q' = aq^b \tag{8.3}$$

Em que a e b são constantes dadas por (8.4) e (8.5)

$$b = \sqrt{\frac{ln(fCV^2 + 1)}{ln(CV^2 + 1)}} \tag{8.4}$$

$$a = \frac{m}{\sqrt{fCV^2 + 1}} \left[\frac{\sqrt{CV^2 + 1}}{m} \right]^2 \tag{8.5}$$

Tal que CV é o coeficiente de variação da distribuição a ser transformada, m o valor médio, e f é o fator de redução da variância.

8.2 Curva de teor e tonelagem

Após a realização da estimativa é interessante resumir os dados obtidos em gráficos para facilitar a visualização dos resultados. Um dos gráficos mais comuns em mineração é o de teor e tonelagem. O teor de cut-off é determinado como aquele em que a mineração começa a ser rentável. A figura

(8.2) demonstra essa ferramenta. Na curva azul temos o percentual de reservas acima do cut-off determinado, enquanto na linha vermelha temos o valor do teor médio acima daquele cut-off. Dessa forma podemos decidir sob diferentes aspectos econômicos a rentabilidade máxima e mínima para o depósito mineral estimado.

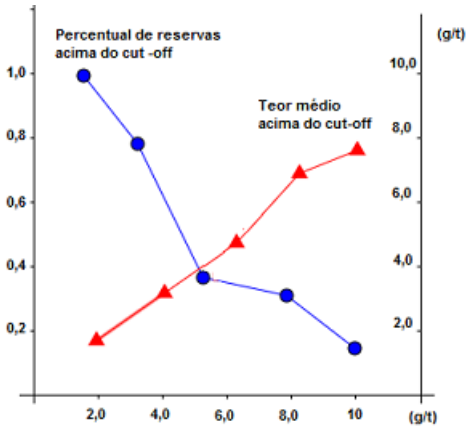


Figura 8.2: Curva teor e tonelagem para um depósito mineral genérico. Na curva azul temos o valor da proporção do depósito para um dado cut-off, enquanto na curva vermelha temos o teor médio acima de um cut-off.

As curvas de teor e tonelagem são usuais em vários estágios da estimativa de depósitos. Durante a exploração mineral ela tem a importância de definir a caracterização inicial de recursos baseados nos dados de amostragem e garantir uma certa visualização de possíveis cenários para a mina. Nesta primeira iniciativa as estimativas ainda não se tornaram uma reserva devido a falta de um estudo de viabilidade. Durante a fase de operação, por exemplo, curvas de teor e tonelagem podem demonstrar possíveis cenários de mudança operacional, indicando as quantidades de material ainda presentes na mina que atendam as condições do beneficiamento.

Curvas de teor e tonelagem podem ser calculadas de diversas formas entre elas temos

- Curvas derivadas de um histograma das amostras
- Curvas derivadas de uma distribuição de probabilidades contínua das amostras
- Curvas derivadas dos blocos estimados
- Curvas baseadas na variância de dispersão dos blocos estimados

8.2.1 Curvas de teor e tonelagem derivadas de histogramas das amostras

Como dito no capítulo 1, as estimativas não podem aumentar ou criar informações acerca do depósito mineral. Tomando esse pressuposto é possível que histogramas desagrupados de amostras possam conter informação necessária para construir curvas de teor e tonelagem representativas, caso o procedimento de amostragem seja coerente. Ao realizar um histograma acumulado podemos encontrar o valor da proporção acima do teor de corte dado por (8.6)

$$P_{V \geq t_c} = 1 - P_{V = t_c} \tag{8.6}$$

Em que $P_{V=t_c}$ é a proporção para um dado teor de corte no histograma acumulado.

Neste caso apenas os valores da classe estarão disponíveis para a inserção no gráfico. Quanto maior for o número de classes do histograma acumulado, maior será o número de pontos a serem plotados. Entre os valores das classes é possível realizar algum tipo de interpolação, lembrando que

a curva é negativa definida, sempre diminuindo com o aumento do teor de corte. Técnicas como polinômios de Hermite são uma forma ideal de se aproximar estas curvas a partir dos dados dos histogramas.

Para o cálculo da média dos valores acima do teor de corte determinado pode-se realizar no histograma a média dos valores das classes pelas suas proporções acima do limite estabelecido. Neste caso teremos um número de pontos também definido pelo número de classes do histograma. A média é neste caso uma função positiva definida, técnicas de interpolação também podem ser utilizadas para se aproximar os valores entre classes.

8.2.2 Curvas de teor e tonelagem a partir de distribuição de probabilidades continuas das amostras

Outra forma adequada de se calcular as curvas de teor e tonelagem a partir de amostras é considerando um ajuste de uma função de densidade de probabilidade para estas. No entanto, isso somente é possível de se fazer quando existe uma distribuição conhecida para o conjunto de dados analisados. Muitas vezes o padrão de proporções das amostras pode apresentar um comportamento não descrito pelas funções mais comuns de densidade de probabilidade.

8.2.3 Curvas de teor e tonelagem baseadas na dispersão dos blocos estimados

Para estudos de viabilidade podemos criar uma curva de teor e tonelagem baseada na dispersão do bloco no volume do depósito. Dessa forma modificamos o histograma das amostras no suporte para o suporte da unidade seletiva de lavra. Utilizando a técnica de transformação afim podemos criar a curva de teor e tonelagem como descrito na seção (8.2.1).

8.2.4 Curvas de teor e tonelagem baseadas na estimativa dos blocos

Ao contrário do procedimento relatado em (8.2.3) as curvas de teor e tonelagem obtidas pela estimativa dos blocos não necessitam de uma mudança de suporte para qualificar as unidades seletivas de lavra. Neste caso ela causa uma suavização nas curvas de teor e tonelagem fazendo com que os teores mais baixos tenham maiores proporções e os teores mais altos menores proporções. O gráfico (8.3) demonstra a relação das curvas de cut-off para a proporção da jazida acima destes.

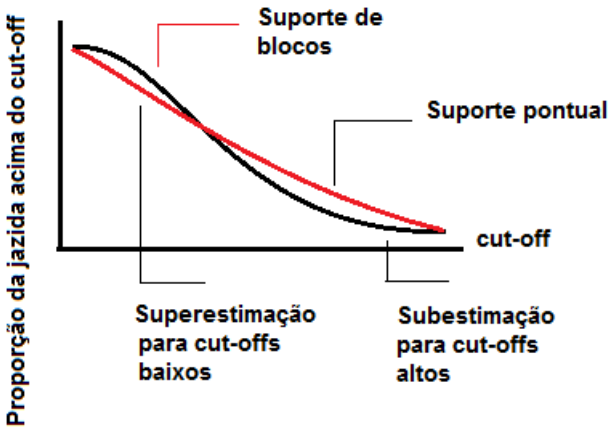
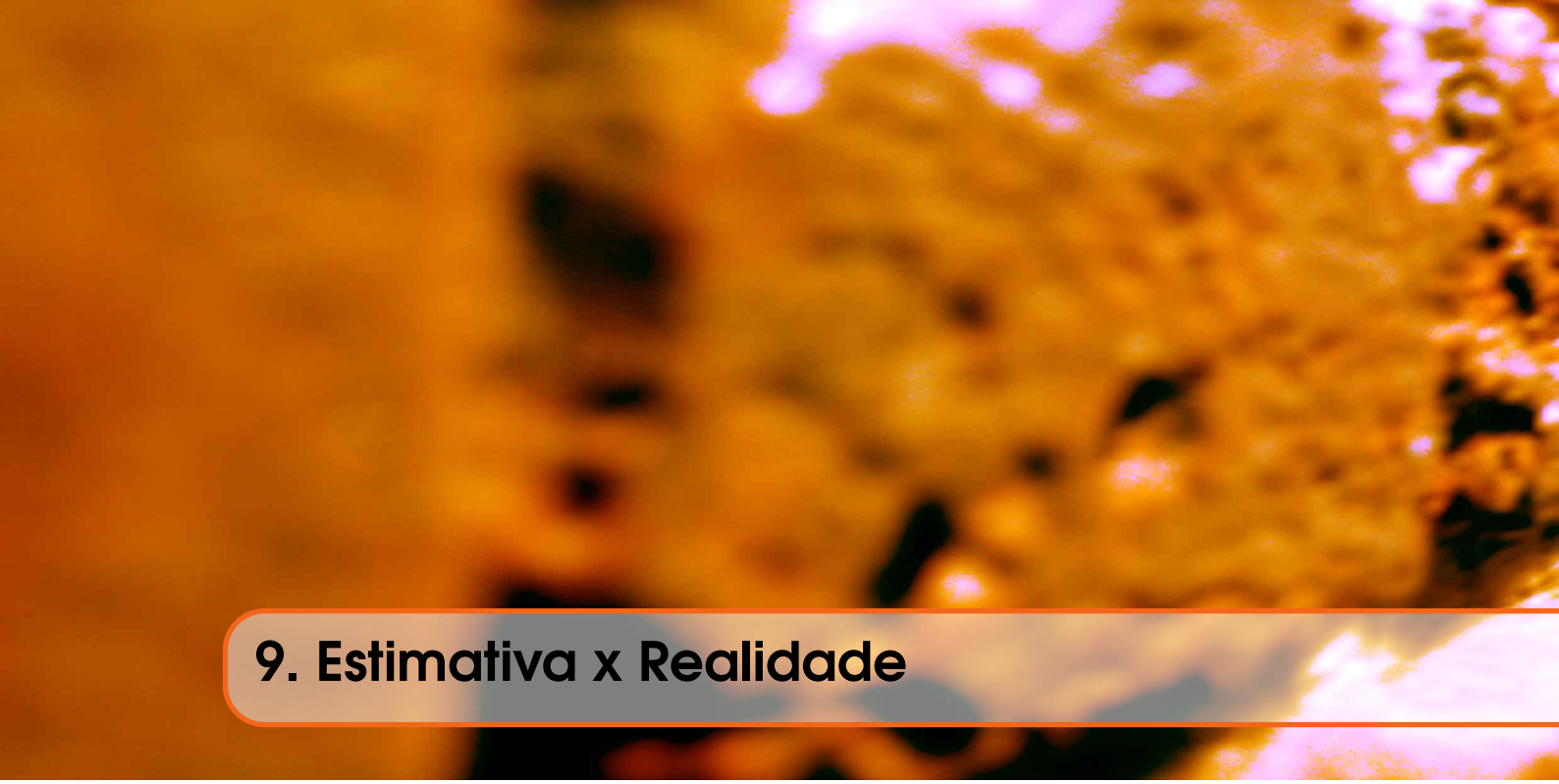


Figura 8.3: Demonstração da suavização da curva teor e tonelagem por mudança de suporte. Valores de cut-off mais baixos recebem maiores proporções, enquanto valores de cut-offs mais altos recebem maiores proporções

8.2.5 Erros associados à determinação da curva de teor-tonelagem

Pequenos erros nas curvas de teor e tonelagem do depósito mineral podem causar grandes variações no retorno do investimento da mineração. Melhores protocolos de amostragem e estimativas mais confiáveis são uma alternativa para se reduzir estas variações. No caso de curvas realizadas a partir de histogramas ou de blocos estimados, o método de interpolação pode interferir na definição da curva. As curvas de teor e tonelagem são sempre enviesadas. Essa diferença dos valores reais e estimados pode ser reduzida com a mudança de suporte. Curvas realizadas com blocos estimados são sempre preferíveis à curvas realizadas de valores pontuais.



9. Estimativa x Realidade

9.1 Introdução

A validação de qualquer estimativa somente pode ser feita compararando os dados estimados com os de produção, comumente referenciado na literatura como estudo de reconciliação. Essa metodologia é aplicada como uma rotina na mineração mas é raramente disponibilizada ao público ou ao meio acadêmico. Diferenças nos teores de elementos metálicos na mineração podem ser causa de quebra de contratos causando prejuízos absurdos para uma empresa.

Existem dois tipos de estudos de reconciliação: aqueles baseados em um banco de dados simulado e aqueles realizados diretamente do depósito mineral. Dados simulados podem comparar diversas formas de estimativa com situações idealizadas da realidade. Para simular um depósito mineral em um certo domínio basta definirmos uma continuidade espacial dos dados e uma distribuição de probabilidades dos dados. Para aproximar os dados simulados da realidade podemos adicionar amostras em suportes já definidos. Criando situações artificiais do depósito podemos verificar o espectro de incerteza que uma estimativa tem sobre a unidade seletiva de lavra.

Outra forma de validação é a comparação de dados de produção com os valores estimados. Antes do material proveniente da lavra, ou também chamado "run of mine", passar pelo processamento mineral, existem amostragens realizadas tanto na bancada como na entrada da usina.

A essência da reconciliação de teores na mineração está em determinar a a variância entre os valores planejados e os de fato obtidos. Existem uma série de técnicas adotadas pelas empresas de mineração envolvendo os estudos de reconciliação, entre eles o controle de teores e de produção, uso de indicadores de performance, reconciliação de recursos e reservas e uso de fatores (mine call factors).

9.1.1 Controle de teores do minério

O controle de teores do minério pode ser visto sob três perspectivas: temporal, espacial e física.

Em relação ao controle temporal, temos os valores diretamente retirados da usina de beneficiamento ou da produção condicionados a um sequenciamento de produção. As diferenças entre os valores estimados e realizados depende o do suporte temporal considerado. Variações mais abruptas tendem a corresponder à tempos pequenos, tais como semanas ou meses, enquanto o planejamento

a longo prazo tende a possuir menores variações.

Sob a perpectiva espacial temos as diferenças entre os recursos planejados e realizados. Por questões operacionais nem sempre a topografia ou a geometria dos stopes são idênticas o que faz o suporte estimado diferente do suporte realizado. Afim de uma comparação é necessário antes de tudo mudar o suporte estimado para o suporte realizado.

Considerando o controle físicos necessitamos que o controle das reservas estimadas estejam coerentes com a mineralização e as densidades prescritas no planejamento, perdas e diluições. A caracterização física depende de uma análise mais profunda, determinando os litotipos e as incertezas de massa e densidade.

9.1.2 Uso de fatores de comparação - forma clássica

O uso de fatores de comparação, geralmente chamados de "Mine Call Factors"tem uso extensivo na indústria e são calculados separadamente dos modelos estimados e o controle diário de teores. A informação necessária para calcular esses fatores são tonelagens, teores do planejamento a longo prazo (modelo de blocos), do planejamento de curto prazo e pelo modelo de controle dos teores. Podemos definir quatro fatores de eficiência em que (9.1) demonstra a eficiência do planejamento a longo prazo:

$$F_1 = \frac{\text{Planejado a curto prazo}}{\text{Planejado a longo prazo}} \tag{9.1}$$

A equação (9.2) demonstra o fator de eficiência para o planejamento de curto prazo

$$F_2 = \frac{\text{Modelo de controle dos teores}}{\text{Planejado a curto prazo}} \tag{9.2}$$

A equação (9.3) demonstra a eficiência da informação passada pela mina

$$F_3 = \frac{\text{Reportado pela mina}}{\text{Modelo de controle dos teores}} \tag{9.3}$$

A equação (9.4) demonstra a eficiência da informação passada pela usina

$$F_4 = \frac{\text{Recebido pela usina}}{\text{Reportado pela mina}} \tag{9.4}$$

Esses fatores levam ao cálculo de de alguns indicadores de performance tais como a precisão do planejamento de longo-prazo (long-term model) (9.5)

$$F_{LTM} = F_1 F_2 F_3 F_4 \frac{\text{Recebido pela usina}}{\text{Planejado a longo prazo}} \tag{9.5}$$

Ou o indicador do planejamento de curto prazo (short-term model) (9.6)

$$F_{STM} = F_2 F_3 F_4 \frac{\text{Recebido pela usina}}{\text{Planejado a curto prazo}} \tag{9.6}$$

A utilização de fatores de performace na mineração sempre deve ser acompanhada da escala de tempo adequada. Para reconciliações a curto prazo é aceitável fazer a reconciliação para valores mensais enquanto para longo-prazo é de se esperar reconciliações de seis meses a um ano.

9.1.3 Uso de fatores de comparação - forma probabilística

O modelo de Parhizkar é geralmente utilizado para realizar a reconciliação da mina baseada nos fatores mais importantes de incerteza na mineração, incluindo a variabilidade inerente, a incerteza estatística e a incerteza sistemática.

A variabilidade inerente é geralmente representada pelo efeito pepita, utilizada nos métodos de estimativa. O modelo de correção geralmente é definido por:

$$G_a = C_r C_s G_e \tag{9.7}$$

Em que G_a e G_e representam os teores medidos e estimados respectivamente. C_r e C_s representam os fatores de correção para os erros estatísticos aleatórios e sistemáticos. Ou seja, C_r representa a correção da variabilidade das amostras e C_s representa a correção do viés.

Podemos obter então o coeficiente de variação para um valor medido de teor como sendo (9.8)

$$CV_{G_a} \simeq \sqrt{\frac{s^2_{G_e}}{\bar{G_e}^2} + \frac{CV^2_{G_e}}{n} + CV^2_{C_1} + CV^2_{C_2}} \tag{9.8}$$

Em que $\frac{s^2_{G_e}}{\bar{G_e}^2}$ representa a variabilidade inerente do fenômeno, dado pela relação da variância dos valores estimados e a média dos valores estimados, $\frac{CV^2_{G_e}}{n}$ representa o erro aleatório da estimativa dado pelo número de amostras n e o coeficiente de variação das estimativas e $CV^2_{C_1}$ e $CV^2_{C_2}$ representa o coeficiente de variação dos fatores de ajuste.

9.1.4 Críticas à geoestatística

A geoestatística lida com a correlação espacial de variáveis aleatórias, a mineração é apenas um dos campos de aplicação deste modelo. Esta é utilizada extensivamente na determinação das estimativas de recurso/reserva, na simulação de depósitos minerais e como ferramentas de auxílio no planejamento e no beneficiamento mineral. Diferentemente dos métodos clássicos o ganho de informação com a krigagem é sem dúvida incomparável. No entanto, é de se esperar que como um modelo, tenha suas próprias falhas. Porventura, a geoestatística é o melhor conjunto de soluções possíveis para a estimativa e simulação de depósitos minerais e não há modelo equiparável na atualidade. Espera-se que com o desenvolvimento de novas metodologias científicas, novas ideias e tendências sobreponham como uma alternativa mais robusta para a solução de problemas na mineração.

Uma das primeiras questões a ser criticada é a descrição da continuidade espacial do depósito mineral. Nos casos mais simples temos a anisotropia definida por um elipsoide, com eixos definidos em uma forma geométrica simples. A correlação espacial de depósitos minerais é naturalmente mais errática e diferente de uma forma geométrica definida. Algumas alternativas propostas atualmente envolvem o desenvolvimentos de mapas de covariâncias, tais que os seus valores sejam tomados diretamente por uma matriz de dados, e não por um modelo geométrico aproximado.

Outra questão a ser criticada é o fato de que a geoestatística é necessariamente fundamentada no uso de estimadores lineares da variável aleatória. Por mais que existam metodologias não-lineares, estas geralmente levam à transformação de distribuições originais das amostras. Isto em certos casos pode acarretar em perda de sensibilidade das distribuições e necessita de valores de correção na transformação dos dados. Algumas metodologias novas, tais como simulação multi-ponto, tendem a evitar o uso de transformações nas distribuições amostrais.

A utilização adequada dos métodos geoestatísticos geralmente é custosa, mesmo que esta beneficie na segurança e na qualidade das avaliações do depósito mineral. A formação de um

geoestatístico treinado requere um maior nível de educação, sendo a mão-de-obra disponibilizada para isso um pouco mais restrita. Os profissionais deste ramo geralmente precisam além da prática cotidiana do método, um alicerce nos conhecimentos básicos de matemática, estatística, programação e geologia. A aplicação adequada da geoestatística envolve não somente o conhecimento na disciplina, mas o reconhecimento e vivência do depósito mineral.

Os procedimentos geoestatísticos são custosos quanto o tempo e demanda computacional. Em alguns casos como estimativas de recursos petrolíferos, as simulações podem durar até mesmo semanas. Em alguns casos o modelo de blocos estimados pode possuir tamanho de memória da ordem de GB. O processamento de dados é volumoso, sendo necessários algoritmos cada vez mais eficientes para lidar com o problema.

A. Geoestatística multivariada

Em muitos casos os problemas relacionados com a estimativa de uma variável de interesse estão associados com duas ou mais variáveis. A geoestatística multivariada é o conjunto de técnicas que permite avaliar concomitantemente mais de uma variável de forma a criar estimativas que incorporem informações diferentes, mas correlacionadas.

Quando realizamos estimativas de variáveis aleatórias diferentes utilizando krigagem ordinária, tal como teores de um dado minério, ocorre a presença de erros de fechamento. Ou seja, se um minério contendo apenas ferro e quartzo em proporções de 60% e 40% não há garantias que as krigagens individuais permanençam com estas proporções. No entanto, utilizando a cokrigagem, por exemplo, conseguimos estimar mantendo as proporções individuais de cada elemento no minério.

Outra utilização da geoestatística multivariada é a incorporação de amostras com suporte diferenciado. Em muitos casos nas campanhas de pesquisa coexistem amostras retiradas por métodos diferenciados tais como sondagem diamantada e pó de perfuratriz. Essas amostras não podem ser utilizadas juntamente pois apresentam precisões e qualidades diferentes e volumes também diferenciados. Utilizando a geoestatística multivariada podemos tratar uma outra informação como uma variável secundária e acrescentar informação que pode qualificar melhor nossa variável de interesse.

Em alguns casos a geoestatística multivariada tem valor muito mais preponderante do que a geoestatística univariada. Em poços de petróleo, em que as informações primárias são escassas, a geofísica de reflexão tem um papel muito mais importante na incorporação da informação na estimativa.

A geoestatística multivariada, tal como a geoestatística convencional, tem os mesmos objetivos principais, mas que, no entanto, se caracterizam pela utilização de múltiplas entradas do modelo. A descrição, interpretação e estimativa são realizadas de forma muito mais complexa e interdependente. Como todo modelo, naturalmente ela não envolverá toda a gama possível de variações e situações encontradas, pois é de fato uma simplificação da realidade. Ao adotarmos a geoestatística convencional, colocamos sob julgamento uma variável objetivo independente de qualquer outro fator, que caracterizará por todo o processo de decisão. A krigagem ordinária, por exemplo, tem

como variável independente as amostras situadas em cada local, e como resposta o valor médio desta variável em um ponto considerado. Mas nada indica que esta variável dependente é apenas condicionada a uma única variável. Tomemos como exemplo um problema físico, demonstrado em (A.1), em que o objetivo é determinar o ponto de parada de uma bola de canhão. Se desconsiderarmos o efeito da resistência do ar, a posição da bola será apenas dependente da velocidade inicial. O modelo inicial é mais simples, mas não garantirá boa precisão, mas se incorporarmos uma variável correlata tal como a resistência do ar, o modelo se tornará mais fiel e semelhante com a realidade. De fato nunca haverá modelo que consiga se acercar de todas as possibilidades de variação, mas cada modelo mais robusto caracterizará melhor as incertezas do problema.

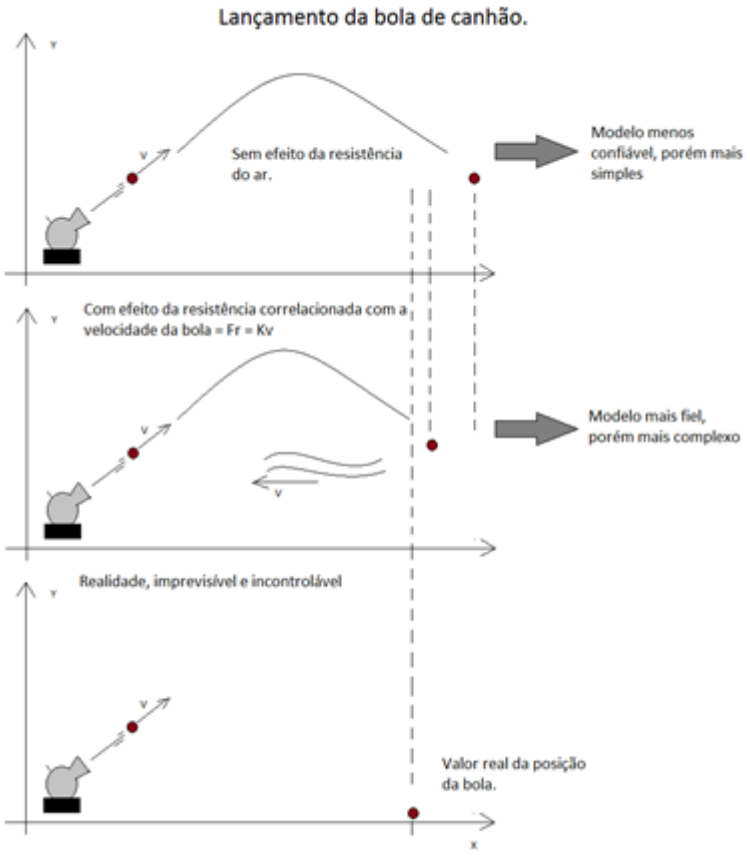


Figura A.1: Modelos físicos diferentes produzindo diferentes resultados. Incerteza sempre dependente do modelo escolhido

A.1 Modelos multivariados

Dentre as metodologias mais comuns de geoestatística multivariada temos:

- Krigagem simples com médias locais variáveis
- Krigagem com deriva externa
- Cokrigagem ordinária
- Cokrigagem colocada

Existem vários outros modelos geoestatísticos multivariados disponibilizados. Para maiores informações procurar "Geoestatistical for Natural Resources Evaluation- Pierre Goovaerts.

A.1.1 Krigagem simples com médias locais variáveis

Como notação para a geoestatística multivariada neste livro mudaremos a notação geralmente utilizada de $Z(x_i)$ para indicar uma variável aleatória no ponto i e mudaremos para $Z_j(x_i)$ tal que j é o índice da variável para o ponto i no espaço.

Lembrando do estimador da krigagem simples tínhamos a equação (A.1) representando a estimativa em um ponto desconhecido:

$$Z^*(x_0) = \sum_{i=0}^{n} \lambda_i \left(Z(x_i) - m\right) + m \tag{A.1}$$

Segundo a hipótese de estacionaridade de segunda ordem o valor de m não depende da posição no espaço sendo um valor constante ao longo de todo o domínio. Podemos utilizar a informação secundária para inferir o valor da média m no ponto desconhecido segundo uma regressão linear. Logo temos a equação da krigagem simples com médias locais variáveis descritas por (A.2):

$$Z_j^*(x_0) = \sum_{i=0}^{n} \lambda_i \left(Z_j(x_i) - msk\right) + msk \tag{A.2}$$

Em que msk é a média regredida entre uma variável j de interesse e uma outra variável qualquer.

A.1.2 Krigagem com deriva externa

Na krigagem com deriva externa não estamos interessados em substituir as médias locais por uma estimativa obtida por regressão linear. Na verdade, neste caso, estamos interessados apenas no modelo a ser utilizado para estas médias.

Geralmente o modelo utilizado para calcular a tendência da função aleatória são polinômios de graus diferenciados. Neste caso a forma mais simples é um modelo linear tal que temos o valor médio igual a $m(x_i) = AZ_2(x_i) + B$, sendo os coeficientes A e B implicitamente calculados pela matriz de krigagem e Z_2 é a variável aleatória secundária. Neste caso o polinômio pode ser caracterizado como uma soma de funções tal que $\sum_{j=0}^{p} a_j f y_j$ sendo p o grau máximo do polinômio e as funções fy sendo os expoentes das variáveis, ou seja $a_1 y_1 + a_2 y_2^2 + \ldots + a_n y_n^n$.

Logo o sistema de krigagem simples pode ser determinado por (A.3):

$$Z_j(x_0) = \sum_{j=0}^{p} a_j f y_j(x_0) + \sum_{i=0}^{n} \lambda_i \left[Z_j(x_i) - \sum_{j=0}^{p} a_j f y_j(x_i) \right] \tag{A.3}$$

Em que a_j são os coeficientes constantes do problema e $f y_j(x_i)$ é o expoente do polinômio no ponto i considerado. Podemos então simplificar a equação acima separando apenas os coeficientes do polinômio, logo podemos ter a equação

$$Z_j(x_0) = \sum_{i=0}^{n} \lambda_i Z_j(x_i) + \sum_{j=0}^{p} a_j \left[f y_j(x_0) - \sum_{i=0}^{n} \lambda_i f y_j(x_i) \right] \tag{A.4}$$

Impondo a restrição que o valor da variável secundária no ponto estimado deve ser igual à uma combinação linear dos valores da variável secundária na região mais próxima temos uma resolução não enviesada do problema tal que:

$$\sum_{j=0}^{p} a_j \left[f y(x_0) - \sum_{i=0}^{n} \lambda_i f y(x_i) \right] = 0 \tag{A.5}$$

Temos então que:

$$fy(x_0) = \sum_{i=0}^{n} \lambda_i fy(x_i) \tag{A.6}$$

Logo o sistema de krigagem nada mais é do que similarmente um sistema de krigagem simples com a restrição dada pela equação (A.6). Para utilizar, no entanto, essa metodologia precisamos ter a variável secundária medida extensivamente. Isso significa que em cada local que estimarmos o valor da variável primária é necessário haver uma medida da variável secundária no ponto a ser estimado e nos pontos utilizados para a estimativa.

A matriz de krigagem fica então modificada para (A.7):

$$\begin{pmatrix} Cov(Y_1,Y_1) & Cov(Y_1,Y_2) & ... & Cov(Y_1,Y_n) & fy(x_1) \\ Cov(Y_2,Y_1) & Cov(Y_2,Y_2) & ... & Cov(Y_2,Y_n) & fy(x_2) \\ ... & ... & ... & ... & ... \\ Cov(Y_n,Y_1) & Cov(Y_n,Y_2) & ... & Cov(Y_n,Y_n) & fy(x_n) \\ 1 & 1 & ... & 1 & 0 \\ fy(x_1) & fy(x_2) & ... & fy(x_n) & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} Cov(Y_0,Y_1) \\ Cov(Y_0,Y_2) \\ ... \\ Cov(Y_0,Y_n) \\ 1 \\ fy(x_0) \end{pmatrix} \tag{A.7}$$

A.1.3 Cokrigagem

Diferentemente da krigagem a cokrigagem utiliza diversas variáveis aleatórias em uma combinação linear de forma a produzir a melhor solução no ponto estimado. A equação (A.8) demonstra como o ponto $Z(x_0)$ pode ser estimado a partir de uma combinação de variáveis aleatórias j:

$$Z_j(x_0) = \sum_{j=0}^{p} \sum_{i=0}^{n} \lambda_i^j Z_j(x_i) \forall j \tag{A.8}$$

Podemos então determinar a variância de extensão da cokrigagem como demonstrado na equação (A.9)

$$\sigma_{ext}^2 = E\left(Z_j(x_0) - \sum_{j=0}^{p} \sum_{i=0}^{n} \lambda_i^j Z_j(x_i) \right)^2 \tag{A.9}$$

Para encontrarmos a matriz de krigagem devemos realizar a expansão da equação (A.9) anterior, tomar o valor esperado de cada termo encontrando as covariâncias e realizar a derivada parcial em relação a cada índice i e j, sendo i o número da amostra e j o número da variável considerada. Observamos na demonstração abaixo que :

Demonstração. $E\left(Z_j(x_0) - \sum_{j=0}^{p}\sum_{i=0}^{n} \lambda_i^j Z_j(x_i)\right)^2$
$E\left((Z_j(x_0))^2 - 2\sum_{j=0}^{p}\sum_{i=0}^{n} \lambda_i^j Z_j(x_i) Z_j^*(x_0) + \sum_{j=0}^{p}\sum_{i=0}^{n}\sum_{j'=0}^{p}\sum_{i'=0}^{n} \lambda_i^j \lambda_{i'}^{j'} Z_j(x_i) Z_{j'}(x_{i'})\right)$
Tomando a esperança matemática de cada parcela temos
$Cov(Z_j(x_0), Z_j(x_0)) - 2\sum_{j=0}^{p}\sum_{i=0}^{n} \lambda_i^j Cov(Z_j(x_i), Z_j^*(x_0)) +$
$\sum_{j=0}^{p}\sum_{i=0}^{n}\sum_{j'=0}^{p}\sum_{i'=0}^{n} \lambda_i^j \lambda_{i'}^{j'} Cov(Z_j(x_i), Z_{j'}(x_{i'}))$
Tomando a derivada parcial em relação a cada $\lambda_i^j \forall i, j$ temos que:
$Cov(Z_j(x_i), Z_j(x_0)) = \sum_{j'=0}^{p}\sum_{i'=0}^{n} \lambda_{i'}^{j'} Cov(Z_j(x_i), Z_{j'}(x_{i'})) \forall i, j$ ■

Esse sistema de krigagem tende a aumentar cada vez mais com a incorporação de mais variáveis secundárias.A figura (A.2) demonstra um exemplo gráfico das partições da matriz de cokrigagem.Neste caso, diferentemente da matriz de krigagem que temos apenas as covariâncias diretas entre ponto a ponto, temos também as variâncias cruzadas.

Figura A.2: Demonstração da matriz de cokrigagem para duas variáveis. Diferentes cores identificam as componentes da matriz

Como demonstrado no capítulo de variografia os covariogramas cruzados não necessariamente são funções pares, e como consequência simétricas. Efeitos de delay podem prejudicar na modelagem de covariogramas sendo a opção de variogramas cruzados a melhor alternativa para a utilização de um modelo linear de corregionalização.

A utilização da cokrigagem não requer que os dados estejam colocados tal como na krigagem com deriva externa. No entanto é necessário determinar um modelo linear de corregionalização de forma a ser capaz a utilização do método. Isso torna a metodologia muito trabalhosa e nem sempre adotada pela maioria dos modeladores exigindo simplificações tais como a utilização de modelos markovianos.

A.1.4 Influência dos dados secundários

A influência dos dados secundários na estimativa da variável primária depende dos seguintes fatores:

- A correlação entre a variável primária e a secundária
- A forma da continuidade espacial entre as variáveis
- A configuração espacial entre as variáveis primárias e secundárias
- a densidade amostral de cada variável

Nota-se que a variável secundária tende a ter maior importância quanto maior for o coeficiente de correlação e menor o efeito pepita relativo entre a variável secundária e a primária. Quanto maior for a qualidade das amostras, ou seja maior acurácia, melhor serão os resultados provenientes da cokrigagem.

A.1.5 Condição não tradicional e tradicional da cokrigagem

Sob a condição tradicional de não enviesamento, espera-se que o sistema de resolução das equações incorpore duas condições de contorno tal que a soma dos pesos de cokrigagem da variável primária seja iguais a 1 e das secundárias seja igual a zero. Essa alternativa é feita para que a variável secundária apenas modifique os pesos de krigagem mas não as unidades do valor estimado. A condição não tradicional, no entanto, propõe que a soma das duas condições seja igual a 1.

A.1.6 Cokrigagem Colocada

A cokrigagem colocada é uma simplificação da cokrigagem convencional, ao qual utilizada dados densamente amostrados para o cálculo dos valores estimados. Na cokrigagem colocada apenas os valores da variável secundária no local onde será estimado o valor da variável aleatória é utilizado. A simplificação permite constatar que a influência da variável secundária é proporcional à distância do ponto estimado, logo ao utilizar a variável apenas no local estimado seu valor "blinda"a influência dos valores mais próximos. Essa condição se torna cada vez mais verdadeira se a correlação entre a variável primária e a secundária tende a ser maior.

B. Geoestatística utilizando o software R

B.1 Introdução

A geoestatística é uma ciência que envolve a manipulação de dados, o que torna imprescindível o uso de programação e softwares. A programação em R é uma linguagem aplicada especificamente para análise estatística computacional e geração de gráficos. Além de possuir uma quantidade grande de bibliotecas que podem ser utilizadas facilmente, existe um grande aporte da comunidade no desenvolvimento e manutenção de novas rotinas.

O R é uma linguagem de programação, e por meio de linhas de comando é possível gerar um algoritmo que permita a análise estatística dos dados fornecidos. Para os inciantes na programação, podemos pensar no algoritmo como uma sequência de instruções a ser realizada para o cumprimento de uma tarefa. Programar, nada mais é, que interagir com o computador, e permitir com que ele faça as tarefas de acordo com suas ordens. Imagine que precisemos fabricar um bolo de chocolate. Para realizarmos estas tarefas realizamos os seguintes passos:

Algoritmo para a fabricação de um bolo

1. Compramos os ingredientes
2. Retiramos os vasilhames da dispensa
3. Misturamos a massa
4. Fabricamos a cobertura
5. Assamos o bolo
6. Cobrimos o bolo com a cobertura

Note que para fabricarmos um bolo precisamos seguir a ordem das instruções, pois não podemos misturar a massa antes de comprar os ingredientes, por exemplo. Essa ordem de predecessão é necessária para que a atividade se cumpra.

No entanto, podemos adicionar estruturas neste algoritmo para que ele se adapte a diferentes condições. Imagine que já tenhamos uma quantidade de ingredientes já comprados. Devemos verificar na dispensa se existe este ingrediente primeiro. Isso pode ser realizado como uma estrutura condicional. O algoritmo se transformaria em:

Algoritmo para a fabricação de um bolo

1. Se ingredientes estão na dispensa
   Não comprar ingredientes
2. Senão
   comprar ingredientes
3. Retiramos os vasilhames da dispensa
4. Misturamos a massa
5. Fabricamos a cobertura
6. Assamos o bolo
7. Cobrimos o bolo com a cobertura

Muita vezes também é necessário realizar tarefas repetitivas, e se torna necessário resumir um número grande de instruções. Neste casso utilizamos estruturas de repetição. Se quiséssemos montar uma fábrica de bolo, poderíamos realizar o seguinte algoritmo:

Algoritmo para a fabricação de um bolo

1. Compramos os ingredientes
2. Retiramos os vasilhames da dispensa
3. Enquanto houver ingredientes faça
   Misturamos a massa
   Fabricamos a cobertura
   Assamos o bolo
   Cobrimos o bolo com a cobertura

Apesar de simples, a fabricação de um bolo ilustra de forma intuitiva o que significa um algoritmo. Para fins de comunicação com uma máquina, se torna necessário o uso de uma linguagem de programação, que permitirá "falar"as instruções para o computador de forma efetiva. No entanto, computadores comunicam apenas com valores binários de 0 e 1. Quanto mais próxima é uma linguagem em comunicar com o computador neste patamar, chamamos esta linguagem de baixo nível. No entanto, quanto mais a linguagem de computação for próxima da linguagem convencional que nós humanos utilizamos, chamamos esta linguagem de alto nível. O R é uma linguagem de alto nível que permite comunicarmos com o computador a partir de instruções praticamente como a escrita em inglês.

Nas próximas seções verificaremos como utilizar um algoritmo e a linguagem R, e como aplicar as funções necessárias para realizar geoestatística em um depósito mineral simples. Utilizaremos o depósito fictício Walker Lake, demonstrado no livro dos professor Issac e Srivastava [8]. Este depósito foi construído a partir de medidas topográficas de uma região em Nevada, no Canadá. As variáveis trabalhadas correspondem a medidas imaginárias V e U.

B.2 Instalação do R

Para utilizar o R precisamos inicialmente instalar o pacote do site `https://www.r-project.org/`. Para utilizar a linguagem é recomendado o uso de uma IDE (Integrated Development Enviroment) de programação. Recomendamos utilizar o RStudio como plataforma para desenvolver os algoritmos. O software pode ser baixado no site `https://www.rstudio.com/`.

B.3 RStudio

O RStudio é uma IDE (Integrated Development Enviroment) gratuita para análise de algoritmos em R. A figura B.1 demonstra as janelas do aplicativo utilizada para as análises estatísticas. No editor é possível criarmos rotinas, escrevendo todas as instruções desejadas e selecionando as desejadas para serem aplicadas. Na janela de variáveis de ambiente observamos todas as variáveis criadas, seu tipo e valores. No console podemos aplicar instruções individualmente, atuando uma instrução de cada vez. E finalmente na janela de output são demonstrados os gráficos, arquivos gerados e pacotes habilitados pelo programa.

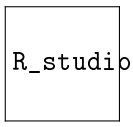


Figura B.1: Demonstração da janela do RStudio. 1 - Editor, 2-Variáveis de ambiente, 3-Console, 4-Output

B.4 Noções preliminares

Muitas vezes desejamos deixar lembretes no nosso código para que futuramente possamos entender melhor as instruções utilizadas. Para realizar comentários no código, utilizamos o símbolo # e em seguida escrevemos o que desejamos na frente do símbolo. Outro importante operador é o de atribuição -> ou =. Este é responsável por associar um valor a uma variável. Uma variável é um objeto capaz de guardar na memória um certo valor, por exemplo um número, um texto, ou até mesmo outro objeto.

```
1
2  nome = "David"                  # Atribui um texto na variavel nome
3  numero = 45.6                   # Atribui um numero na variavel
4  grafico = hist(dados$V)         # Atribui um objeto na variavel
```

Em alguns casos podemos atribuir uma variável a ela mesma. Lembre-se que o operador = não significa igualdade, mas atribuição. O exemplo abaixo demonstra este resultado.

```
1
2  numero = 5              # Atribui 5 na variavel
3  numero = numero + 1        # Adiciona um na variavel 5, retornando 6
```

B.5 O R como uma calculadora

Assim como uma calculadora comum o R realiza cálculos básicos como subtração, adição, subtração, soma, etc. Deve-se lembrar sempre da predecessão dos operadores matemáticos no momento de realizar os cálculos, podendo ser utilizados parênteses para definir as relações de cálculo.

```
1
2  2+2 # Soma
3  2-2 # Subtracao
4  2*2 # Multiplicacao
5  2/2 # Divisao
6  2^2 # Exponenciacao
7  7%2 # Resto da divisao
```

```
3*(4-3)^2
```

B.6 Utilizando funções no R

Para utilizar uma função é necessário escrever seu nome e adicionar seus argumentos dentro de parênteses. Usualmente atribui-se o nome do argumento e então é atribuído seu valor. Podemos passar argumentos sem seus nomes também, e dessa forma a predecessão de cada um destes é atribuído segundo a ordem estabelecida pela função. Abaixo encontramos algumas funções comuns do R.

```

log(3)          # logaritmo natural de 3
sqrt(45)        # raiz quadrada de 45
factorial(4)    # fatorial de 4 , 4!
abs(5-3)        # valor absoluto de 2
```

Listing B.1: Operacoes matemáticas convencionais utilizando o R

B.7 Operadores Relacionais

Operadores relacionais são aqueles utilizados para comparar valores entre números ou expressões. O resultado de um operador relacional é um valor booleano que indica se a expressão é verdadeira ou falsa. Os operadores utilizados no R são demonstrado na tabela abaixo. É importante lembrar que o operador = é associado à atribuição de uma variável ao contrário do operador == que representa igualdade:

Operador	Relação
==	Igualdade
!=	Diferente
>	Maior
<	Menor
>=	Maior e igual
<=	Menor e igual

Tabela B.1: Operadores relacionais no R

Abaixo é apresentado alguns resultados dos operadores relacionais para o R.

```
3 == 5          # Retorna FALSO
5 > 3           # Retorna VERDADEIRO
-2 < 7          # Retorna VERDADEIRO
abs(-5+3)== 2   # Retorna VERDADEIRO
4 != 4          # Retorna FALSO
```

Listing B.2: Exemplo de operadores relacionais no R

B.8 Operadores Lógicos no R

A utilização de operadores booleanos na programação é algo muito comum quando precisamos avaliar relações múltiplas. A lógica matemática é o fundamento principal de um operador lógico que retorna um valor booleano (Verdadeiro ou Falos) a partir de um conjunto de relações. As operações mais comuns são o E, simbolizado por & e o OU, simbolizado por ||.

O operador E retorna valor verdadeiro apenas se as duas premissas relacionadas forem verdadeiras. A frase "A ferrari é uma marca de carro, E é muito cara"retorna valor verdadeiro, pois ambas são de fato verdade. A tabela verdade abaixo demonstra o resultado do operador E assumindo os valores de cada uma das premissas utilizadas.

Valor 1	Valor 2	Operação E
Verdadeiro	Verdadeiro	Verdadeiro
Falso	Verdadeiro	Falso
Verdadeiro	Falso	Falso
Falso	Falso	Falso

Tabela B.2: Tabela verdade do operador E

Já o operador OU retorna valor falso apenas se as duas premissas relacionadas forem falsas, retornando verdadeiro em todos os outros casos. A frase "A ferrari não é uma marca de carro, OU é muito cara"retorna valor verdadeiro, pois a marca é muito cara. A tabela verdade abaixo demonstra as relações do operador OU.

Valor 1	Valor 2	Operação OU
Verdadeiro	Verdadeiro	Verdadeiro
Falso	Verdadeiro	Verdadeiro
Verdadeiro	Falso	Verdadeiro
Falso	Falso	Falso

Tabela B.3: Tabela verdade do operador OU

A script abaixo demonstra operações relacionais no R

```
(5>4) & (3<7)          # Retorna VERDADEIRO
(5>4) & (3<2)          # Retorna FALSO
(7==4) || (2>3)        # Retorna FALSO
(3>2) || (2>3)         # Retorna VERDADEIRO
```

Listing B.3: Exemplo de operadores relacionais no R

B.9 Pedindo ajuda no R

Muitas vezes não conhecemos adequadamente o funcionamento de alguma função ou comando. Para procurar ajuda, o R possui a função help() que auxilia na identificação dos argumentos da função, ou simplesmente pode-se colocar o símbolo ? antes da função que se pretende identificar.

```
1
2  help(par) # Ajuda para a funcao par
3  ?par      # Ajuda para a funcao par
```

Listing B.4: Exemplo de ajuda utilizando a função help

B.10 Pacotes do R

Uma das grandes vantagens da utilização do R consiste em sua grande quantidade de pacotes disponíveis de todos os tipos. É necessário no R instalar estes pacotes utilizando o comando install.packages. A opção "dependences"permite com que o pacote instale qualquer outro tipo de dependência necessária para o funcionamento do pacote, assumindo valor verdadeiro = T ou TRUE, ou valor falso = F ou FALSE. Utilizaremos neste livro dois pacotes importantes para a análise dados, o pacote sp, responsável por análises espaciais e o pacote gstat e geoR, responsáveis para realizar a análise geoestatística.

```
1
2  # Comandos para instalacao dos pacotes
3  install.packages(sp, dependences=T)
4  install.packages(gstat, dependences =T)
5  install.packages(geoR, dependences = T)
6
7  # Comandos para carregar os pacotes
8  library(sp)
9  library(geoR)
10 library(gstat)
```

Listing B.5: Código fonte em R para instalação dos pacotes necessários

B.11 Criando vetores

Vetores são objetos capazes de armazenar vários dados. É possível armazenar em vetores tanto variáveis numéricas como também textos, porém, apenas um tipo de variável deve ser adicionado em cada vetor. Para criar um vetor de dados inicia-se com a letra c, colocando os valores na ordem desejada separados de vírgula. O código abaixo demonstra a criação de um vetor de dados.

```
1
2  # Criacao de um vetor de numeros
3  dedos = c(1,2,3,4,5,6,7,8,9,10)
4
5  # Criacao de um vetor de textos
6  aves = c("tucano", "gaivota", "pombo")
```

Listing B.6: Criação de um vetor em R

Algumas operações podem ser realizadas com estes vetores.

```
1
2  # Criacao de um vetor de numeros
3  dedos = c(1,2,3,4,5,6,7,8,9,10)
4
5  max(dedos)          # Retorna o valor maximo do vetor dedos
```

```
6  min(dedos)          # Retorna o valor minimo do vetor dedos
7  sum(dedos)          # Retorna a soma dos itens do vetor dedos
8  length(dedos)       # Retorna o tamanho do vetor dedos
```

Listing B.7: Criação de um vetor em R

Para acessar um valor do vetor podemos utilizar um colchetes para indicar a posição do elemento desejado. Caso deseje retornar o vetor sem um elemento podemos usar índices negativos.

```
1
2  # Criacao de um vetor de numeros
3  vetor= c(5,4,12,11,45,6,7)
4
5  vetor[1]             # Retorna 5
6  vetor[2]             # Retorna 4
7  vetor[-1]            # Retorna 4,12,11,45,6,7
```

Listing B.8: Criação de um vetor em R

Podemos gerar sequências de números também utilizando os dois pontos. Por exemplo, para gerar números de um a dez podemos usar o comando 1:10. Podemos gerar sequências de números também utilizando a função seq(), para isso utilizamos os atributos from, para identificar o número de início, to, para identificar o valor final e by, para identificar o passo de um número para outro. Podemos gerar também repetições com o comando rep(). O resultado de rep(5,4) será c(5,5,5,5).

```
1
2  1:10                          # gera 1 2 3 4 5 6 7 8 9 10
3  seq(from=1, to=10, by=2)       # gera 1 3 5 7 9
4  rep(5,4)                        # gera 5 5 5 5
```

Listing B.9: Criação de um vetor em R

B.12 Condicional

Como visto no exemplo do bolo, podemos indicar condições para o cumprimento de uma determinada tarefa. Ao utilizar o operador IF, conseguimos determinar se instruções serão realizadas ou não de acordo com uma condição. No exemplo abaixo o algoritmo verifica se o valor C é maior que 5, e em seguida o valor de C é demonstrado na tela, caso o contrário é utilizado o comando else, e demonstrado na tela o valor de 5. Para separar uma instrução condicional de outra é utilizado o colchetes.

```
1
2
3  If (C > 5){
4    print(C)
5  }
6  else{
7    print(5)
8  }
```

Listing B.10: Criação de um vetor em R

B.13 Repetições

Commo visto no exemplo do bolo, podemos repetir instruções de acordo com uma ordem. Um dos operadores que pode ser facilmente utilizado para repetições é o for. Outro tipo de repetição é quando utilizamos a estrutura while, em que a repetição ocorre até encontrar uma condição de

parada. Muito cuidado deve-se ter ao utilizar a estrutura while, pois se a repetição não encontrar a condição de parada ela se repetirá infinitamente, consumindo a memória do computador. O exemplo abaixo plota os dez primeiros números de uma sequência fornecida utilizando tanto o comando for como while.

```
1
2  for (i in 1:10){
3    print(i)
4  }
5
6  i = 1
7  while(i <= 10){
8    print(i)
9    i = i + 1
10 }
```

Listing B.11: Criação de um vetor em R

B.14 Concatenação de funções

Em muitos os casos podemos concatenar funções dentro de outras funções, assim como também podemos concatenar operadores dentro de operadores. As condicionais podem ser realizadas uma dentro das outras, assim como as repetições podem ser realizadas uma dentro das outras. Veja os exemplos abaixo

```
1
2  y = sqrt(abs(-16))  # Retorna 4
3
4  if (C> 5){          # Verifica a primeira condicao
5    if (C/2 == 3){      # Verifica a segunda condicao
6      print("ok")
7    }
8  }
```

Listing B.12: Criação de um vetor em R

B.15 DataFrames

Para trabalhar com dados é necessário organização, de forma a acessar os valores de forma rápida e eficiente. O DataFrame é um dos objetos do R responsáveis por organizar estes dados. A tabela abaixo demonstra um DataFrame para o conjunto de dados do Walker Lake. No topo temos o nome de cada coluna (ID, V, U, T), enquanto a esquerda temos o nome de cada linha, representado pelos números 1,2,3. Para acessar uma coluna do DataFrame devemos escrever o nome do dataframe, colocar um sinal de $, em seguida o nome da variável.

	Id	V	U	T
1	1	0.0	NA	2
2	2	0.0	NA	2
3	3	224.4	NA	2

Tabela B.4: Exemplo de DataFrame no R

A biblioteca sp contém em seus bancos de dados internos o depósito do Walker Lake. Para acessar estes dados basta apenas utilizar a função data(walker) e assim estará disponível a variável

walker para uso. Para vizualizarmos alguns dados do banco podemos utilizar a função head() em que é mostrado as primeiras linhas dos dados. Podemos também utilizar a função tail() para verificar os últimos dados do banco. Uma função importante é a função summary, que permite realizar um resumo estatístico dos dados.

```
1
2  # Comandos para instalacao dos pacotes
3  install.packages(sp, dependences=T)
4  install.packages(gstat, dependences =T)
5  install.packages(geoR, dependences = T)
6
7  # Comandos para carregar os pacotes
8  library(sp)
9  library(geoR)
10 library(gstat)
11
12 data(walker)              #Baixa o conjunto de dados do Walker Lake
13 head(walker)              # Observa os primeiros valores do Walker Lake
14 tail(walker)              # Observa os ultimos valores do Walker Lake
15 summary(walker)           # Realiza um sumario estatistico do Walker Lake
```

Listing B.13: Criação de um vetor em R

No entanto, nem sempre é comum trabalharmos com dados já disponibilizados nas bibliotecas. O R possui funções para importação de dados CSV, excel e de banco de dados. Para isso podemos utilizar a função read.table() ou read.csv() para abrirmos um arquivo de texto ou csv. A função file.choose() permite com que uma janela para arquivos seja aberta, facilitando encontrar o endereço do arquivo. O argumento header verifica se o banco de dados possui um cabeçalho e o argumento sep verifica qual separador é utilizado para dividir as colunos no banco de dados. No caso de arquivos csv, o separador é a vírgula.

```
1
2  # Importacao de dados a partir de uma tabela ou arquivo csv
3  dados = read.table(file.choose(), header =TRUE)
4  dados = read.csv(file.choose(), header= TRUE, sep=",")
```

Listing B.14: Criação de um vetor em R

B.16 Mapa de localização

O posicionamento das amostras no mapa é de extrema importância para a análise espacial dos dados. Para realizarmos um mapa de localização das amostras podemos utilizar a biblioteca geoR, transformando o dataframe em um arquivo geodata e em seguida aplicando a função points.

```
1
2  points(as.geodata(walker$V))
```

Listing B.15: Criação de um vetor em R

O resultado do gráfico pode ser demonstrado na figura abaixo. Nota-se que o Walker Lake apresenta uma malha regular amostrada e valores agrupados em corpos específicos, um maior situado no flanco oeste e corpos menores situados no flanco leste.

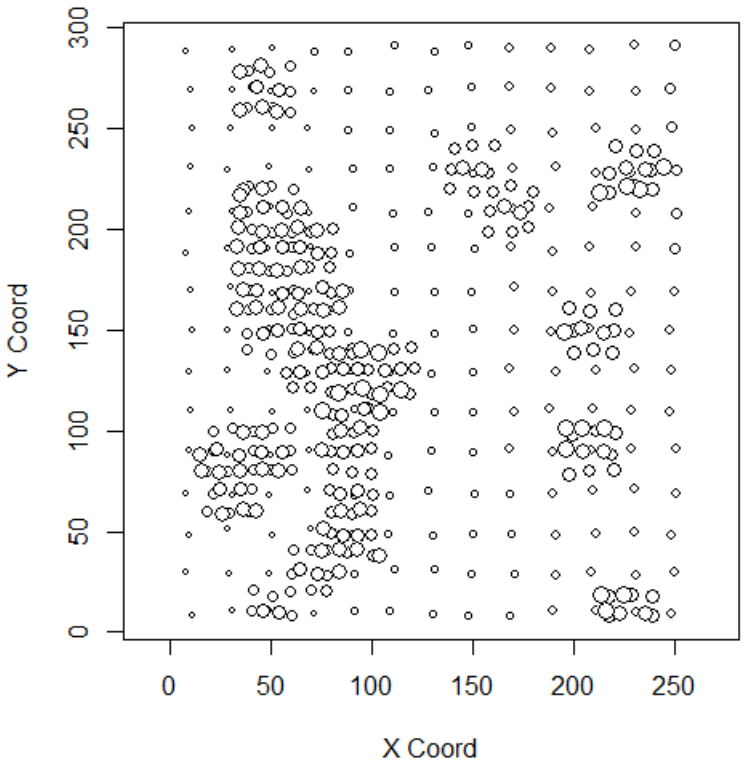


Figura B.2: Mapa de localização das amostras do Walker Lake

Outro gráfico com excelente visualização dos valores das variáveis pode ser realizado com o ssplot(). Mas antes para gerar os dados precisamos associar ao banco de dados do Walker Lake suas coordenadas. Para isso usamos o comando coordinates() e a ele associamos os valores das variáveis X e Y.

```
1
2
3   # ASSOCIAR AS COORDENADAS NO ESPACO
4   coordinates(walker) = c("X", "Y")
5
6   # SSPLOT DA VARIAVEL V
7   spplot(walker, c("V"),scales = list(draw =T))
8
9   # SSPLOT DA VARIAVEL  V E U
10  spplot(walker, c("V","U"),scales = list(draw =T))
```

Listing B.16: Criação de um vetor em R

A figura abaixo demonstra o mapa de intervalos para o depósito do Walker Lake. Os valores da variável V se alteram de 0 para 1528, e podemos notar que as regiões mais ricas se situam dentro do maior corpo do depósito na região oeste.

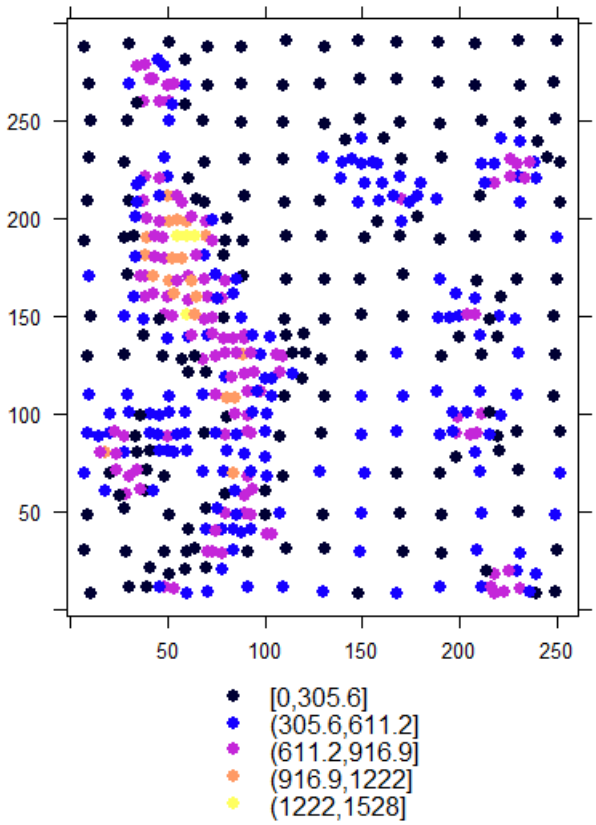


Figura B.3: Mapa de intervalos das amostras do Walker Lake, variável V

A Figura abaixo demonstra as variações para concumitantes para as variáveis V e U, dessa forma conseguimos visualizar a correlaçào entre as variáveis tal como o local onde foram amostradas.

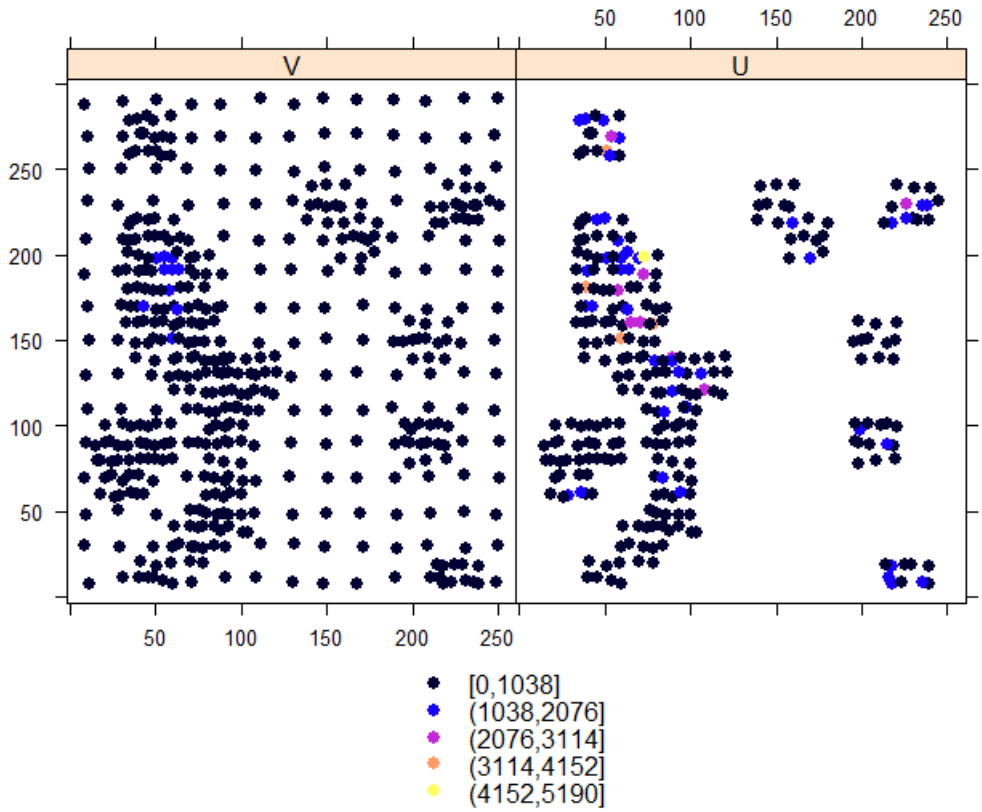


Figura B.4: Mapa de intervalos das amostras do Walker Lake, variável U e V

B.17 Histogramas

Podemos gerar histogramas das variáveis de interesse utilizando a função hist(), sendo o primeiro argumento a variável utilizada na construção do gráfico. O número de classes pode ser selecionado de acordo com o argumento breaks. Os argumentos xlab, ylab e main apenas definem o nome dos eixos plotados no gráfico. O R permite a utilização de múltiplos gráficos na mesma figura, isso pode ser obtido utilizando o comando par, e fornecendo um vetor para o argumento mfrow com o número de linhas e de colunas, respectivamente.

```
1
2
3  # funcao para criar mais de um grafico junto
4  par(mfrow = c(1,2))
5
6  # Adicionar grafico da variavel U
7  hist(walker$U, main= "histograma da variavel U ", breaks =15, xlab= "U", ylab
       = "Frequencia")
8
9  # Adicionar grafico da variavel V
10 hist(walker$V, main= "histograma da variavel V", breaks = 15,  xlab= "V",
       ylab = "Frequencia")
```

Listing B.17: Criação de um vetor em R

A figura abaixo demonstra os histogramas gerados pelo código fonte. Notamos uma alta assimetria na variável U, enquanto a variável V demonstra valores mais espaçados entre si.

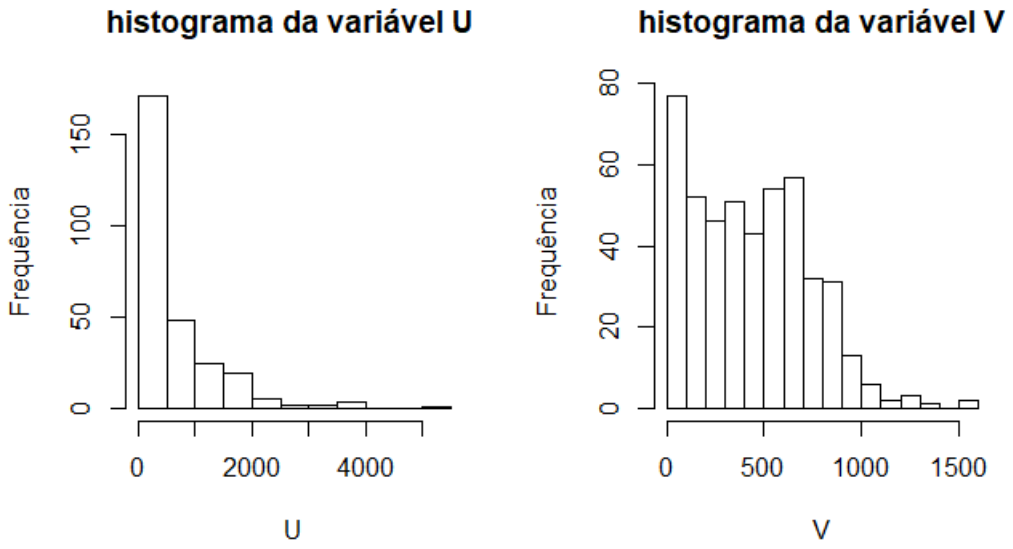


Figura B.5: Histogramas das variáveis U e V do Walker Lake

B.18 Boxplots

Gráficos de caixa, ou também chamados de "boxplot"são uma ferramenta importante para avaliação de valores outliers que podem distorcer as estatísticas. Muito cuidado deve ser tomado na hora do tratamento de valores anômalos. Se as distribuições forem altamente assimétricas o gráfico pode apresentar um número muito grande de valores anômalos falseados, sendo necessário cautela na remoção destes valores.

```
1
2
3  par(mfrow = c(1,2))
4  boxplot(walker$U, main= "Boxplot da variavel U ", ylab= "U")
5  boxplot(walker$V, main= "Boxplot da variavel V",  ylab= "V")
```

Listing B.18: Criação de um vetor em R

A figura abaixo demonstra o gráfico de caixas utilizado para a modelagem matemática do Walker Lake. Notamos que a variável U apresenta um ponto discrepante acima de 5000.

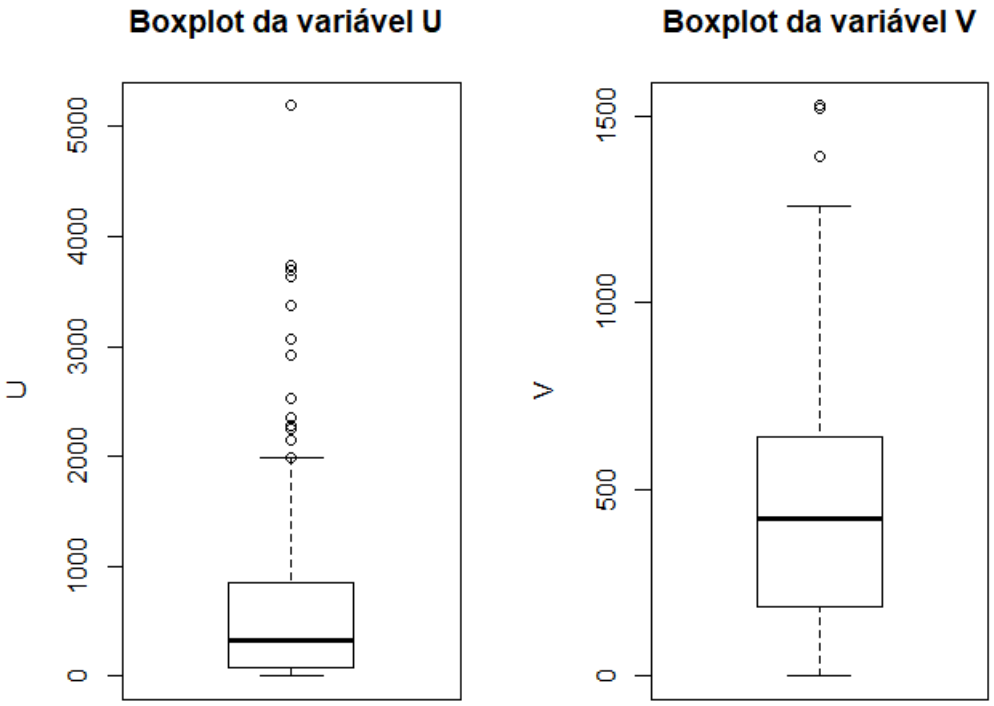


Figura B.6: Boxplot das variáveis U e V do Walker Lake

B.19 Regressão Linear

Para realizarmos um modelo de regressão linear podemos utilizar o comando lm() em que primeiramente informamos a variável Y, e em seguida informamos a variável X separando-a por um sinal de ~. Para obtermos informações sobre os valores da regressão, basta utilizar a função summary, ao qual será informadas várias estatísticas, inclusive o coeficiente de regressão de Pearson. Em seguida para plotarmos o gráfico podemos utilizar a função plot, informando os valores de X e de Y. A reta de regressão pode ser adicionada utilizando o comando abline() e informando como argumento o modelo linear.

```
1
2
3  #Regressao Linear
4  linear = lm(walker$U~walker$V)
5  summary(linear)
6
7  # Plotagem do resultado
8  plot(walker$V, walker$U, xlab="V", ylab="U")
9  abline(linear)
```

Listing B.19: Criação de um vetor em R

O gráfico abaixo representa o modelo de regressão para as variáveis V e U. Um dos pontos acima do valor de 5000 parece demonstrar um valor outlier.

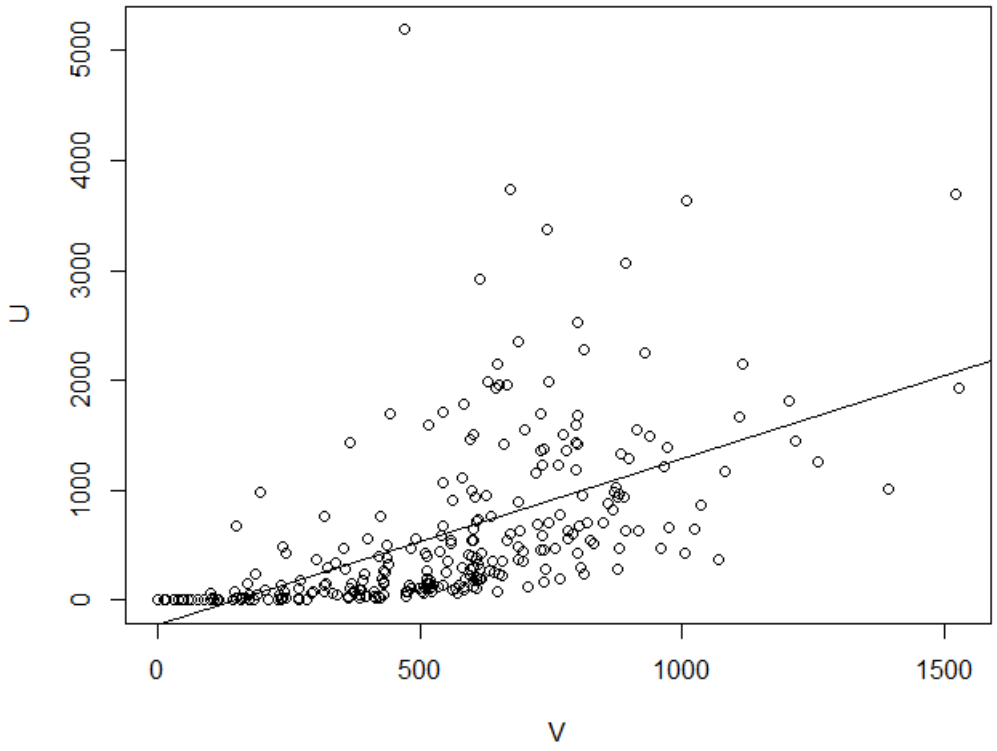


Figura B.7: Regressão linear das variáveis U e V do Walker Lake

B.20 Vizinho mais próximo

Como metodologia para o desagrupamento, optamos por utilizar o vizinho mais próximo para encontrar as estatísticas desagrupadas. Primeiramente precisamos criar um grid, onde será realizada as interpolações. A função makegrid constrói um grid com um tamanho de célula definida pelo argumento cellsize. Para que possamos observar os resultados em um mapa de pixels, precisamos realizar algumas conversões, transformando primeiro em um objeto de pontos e em seguida de pixels. Em seguida podemos realizar a interpolação por vizinhos mais próximos, criando um raster dos dados e um objeto gstat que será utilizado na interpolação. Neste último definimos o número máximo de pontos considerados na interpolação do vizinho mais próximos. Ao atribuirmos nmax = 1 dizemos que cada valor da célula receberá única, e exclusivamente o valor da amostra mais próxima desta.

```
1
2
3   # Criar um grid
4   grid_stat = makegrid(walker,cellsize = 5)
5   grid_stat = SpatialPixels(SpatialPoints(grid_stat))
6
7
8   # Calcular vizinho mais proximo
9
10  ca = raster(walker, res= 1)
11  gs = gstat(NULL, "V", V~1, walker, nmax=1)
```

```
12  nn = interpolate(ca, gs)
13  plot(nn, axes=T)
```

Listing B.20: Criação de um vetor em R

O resultado da interpolação pode ser compartilhado na figura abaixo.

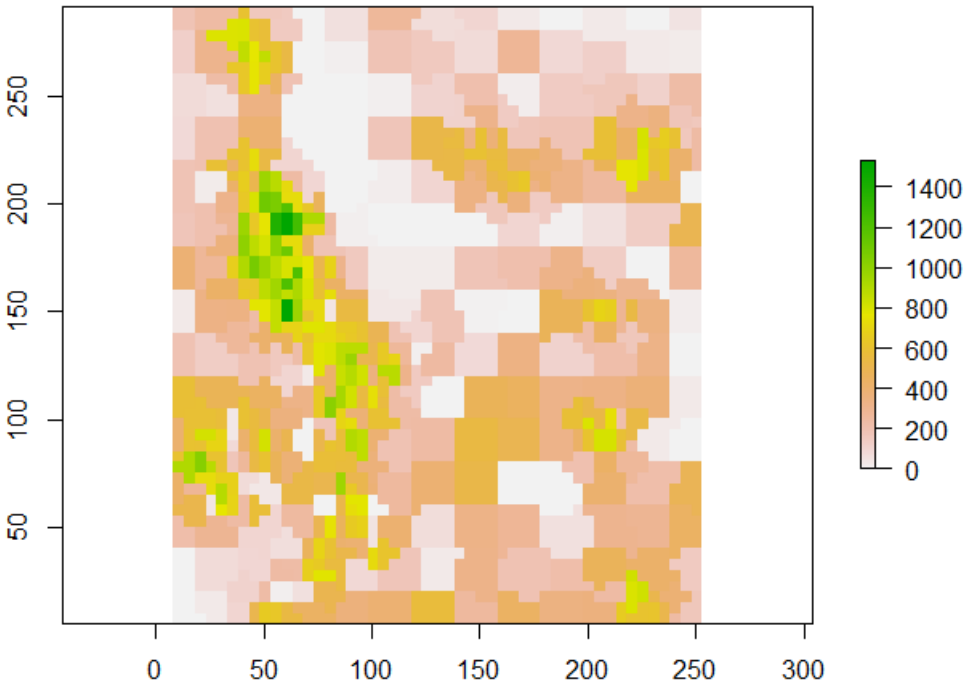


Figura B.8: Vizinho mais próximo Walker Lake - variável V

B.21 Variograma

A variografia é uma das peças fundamentais para a criação de um modelo interpolado utilizando técnicas de geoestatística. Para avaliar a qualidade da dependência espacial dos dados, podemos utilizar uma ferramenta muito conhecida chamada de gráfico de dispersão h. Abaixo vemos o código fonte para a geração do gráfico. Primeiramente fornecemos o valor da variável a ser medida, em seguida o dataframe ao qual ela está contida e por fim o vetor com as distâncias para cada gráfico de dispersão.

```
1
2  # H-scatterplot da variavel V
3  hscat(V~1, walker, (0:9)*5)
```

Listing B.21: Criação de um vetor em R

A figura abaixo demonstra o gráfico de dispersão h para diferentes distâncias. Notamos que a correlação entre as variáveis distanciadas tende a cada vez mais decrescer de acordo com a distância entre os dados.

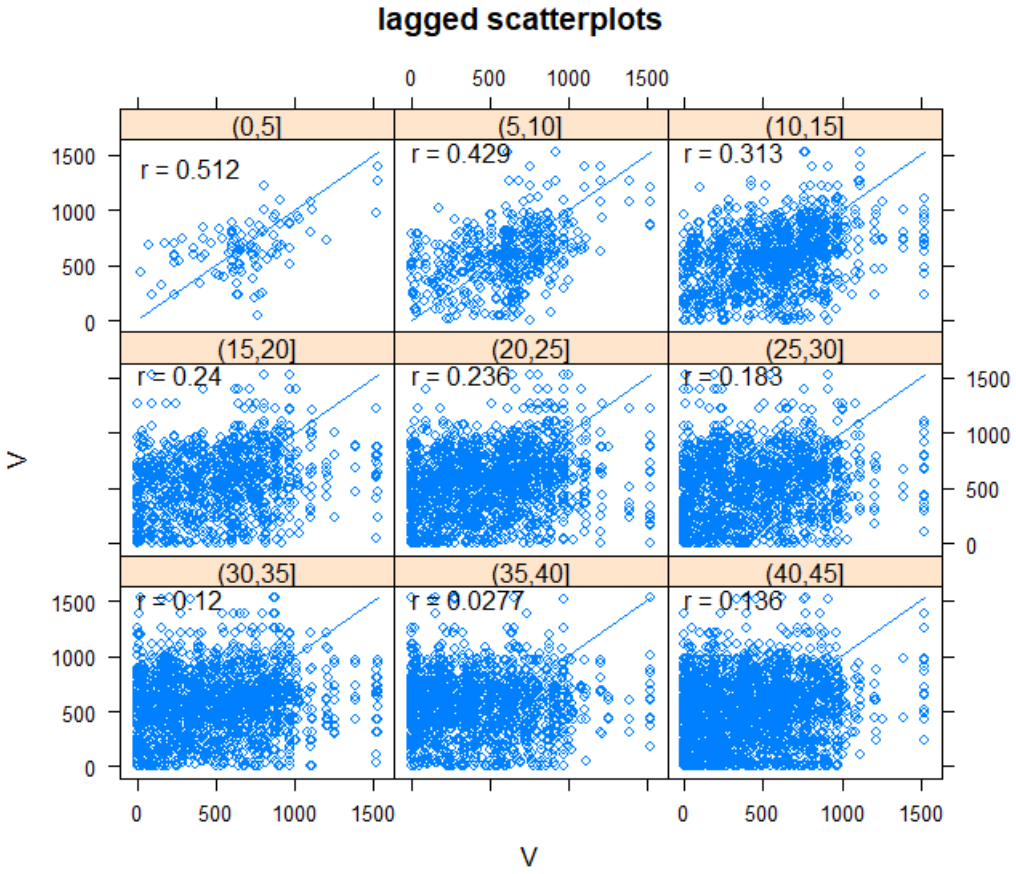


Figura B.9: Hscatterplot para a variável V

Variogramas são a principal ferramenta para a análise de continuidade espacial. Primeiramente precisamos saber a variância da variável modelada V, para que encontremos o valor máximo do patamar. Para isto utiilizamos o comando var(). Em seguida é necessário realizar o variograma experimental dos dados. Para isto utilizamos o comando variogram(). O primeiro argumento fornecido corresponde a variável utilizada. O segundo argumento corresponde ao dataframe considerado. O argumento width corresponde ao lag ou espaçamento utilizado para o cálculo dos variogramas experimentais. O cutoff representa a distância máxima para se calcular o variograma. Finalmente informamos a tolerância horizontal de cada um dos variogramas em graus. Como o problema é bidimensional, as direções de cada um dos variogramas é controlada apenas pelo azimute. O argumento alpha corresponde a uma lista de valores ao qual será calculado o variograma, para identificarmos a direção de máxima continuidade.

```
1
2
3  #Variancia da variavel V
4  var(walker$V)
5
6  #Variogramas experimentais
7
8  v.dir = variogram(V~1, walker, width=10, cutoff= 200, tol.hor=45, alpha =
       (0:7)*22.5 )
```

Listing B.22: Criação de um vetor em R

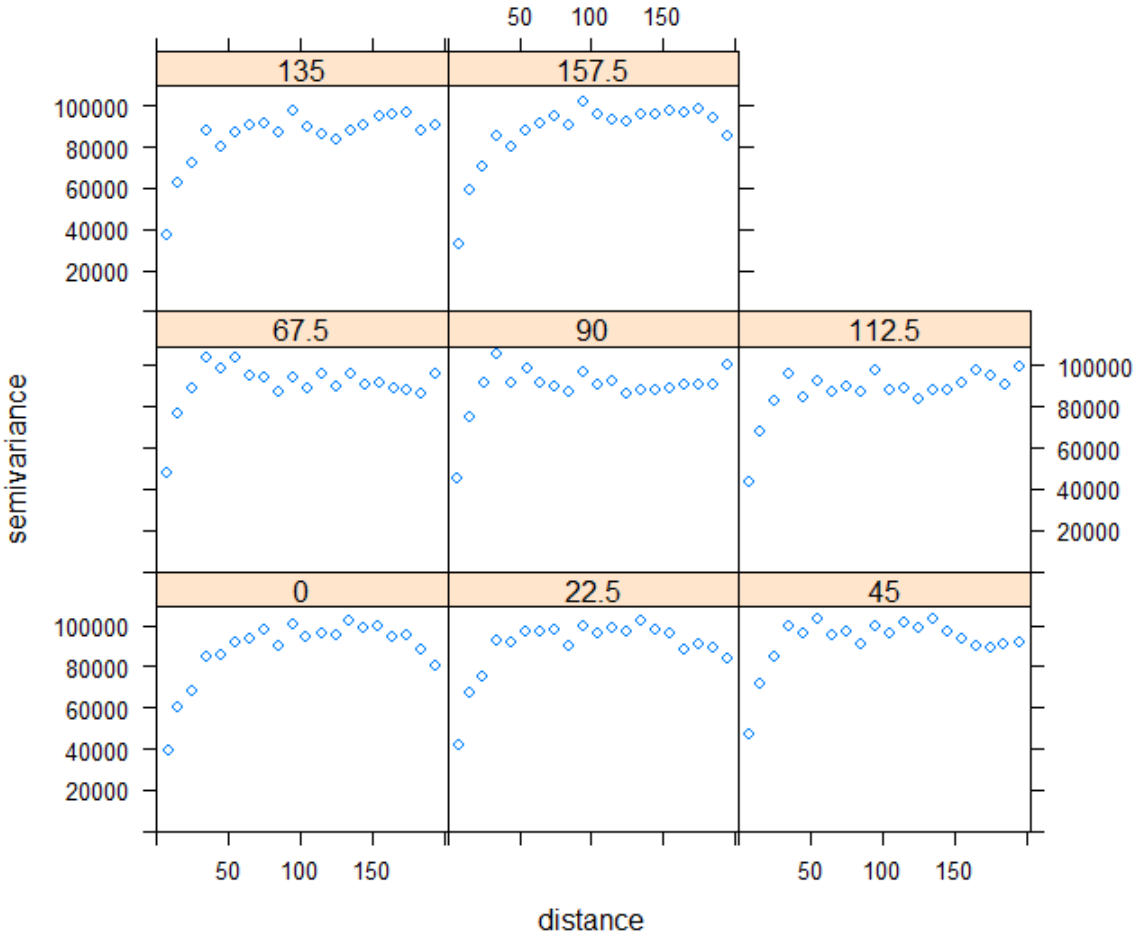


Figura B.10: Variogramas experimentais gerados

A modelagem de variogramas individuais pode ser realizada a partir da biblioteca geoR utilizando o comando eyefit. para isso criamos um objeto do variograma experimental utilizando a função variog() e atribuímos os argummentos relacionados com o variograma, tais como direção (em radianos), tolerância angular (em radianos), o argumento uvec, que representa o vetor com distâncias a serem calculadas e a máxima distância de cáculo. Lembre-se que para utilizar as rotinas do pacote geoR é necessário transformar o banco de dados em um tipo específico chamado de geodata.

```
1
2
3   library(geoR)
4   library(gstat)
5
6   # Baixar os dados do Walker Lake
7   data(walker)
8
9   # Associar coordenadas ao banco de dados
10  coordinates(walker) = c('X', 'Y')
11
12  #Criar variograma experimental utilizando a biblioteca geoR
13  var = variog(as.geodata(walker["V"]), uvec = seq(from=0,to=500,by=10),max.
```

```
    dist = 500, direction= 157.5*pi/180, tolerance = pi/4)

#Fitar manualmente o variograma
ve.eye = eyefit(var)

#Transformar o modelo fitado em um modelo do gstat
ve.fit = as.vgm.variomodel(v.eye[[1]])
```

Listing B.23: Criação de um vetor em R

A figura abaixo demonstra o ajuste do modelo de variograma utilizando a função eyefit. Uma janela é aberta podendo ser selecionado os modelos mais adequados para o ajuste, tal como é possível também selecionar os melhores parâmetros como patamar, alcance e efeito pepita.

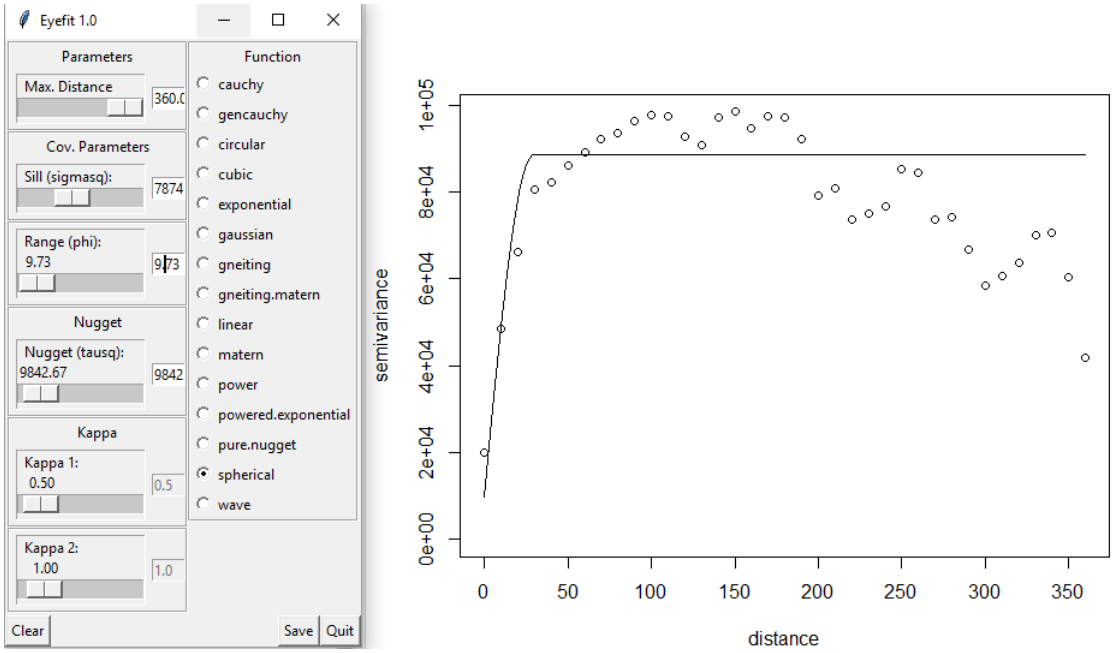


Figura B.11: Ajuste do modelo de variograma utilizando a função eyefit

Para ajustarmos um modelo de variograma podemos utilizar a função vgm() para as diferentes direções e para diferentes estruturas. Diversos são os tipos de modelos aceitados pela função. A tabela abaixo demonstra uma relação dos modelos abordados pela função vgm.

	Forma curta	Forma longa
1	Nug	Nug (nugget)
2	Exp	Exp (exponential)
3	Sph	Sph (spherical)
4	Gau	Gau (gaussian)
5	Exc	Exclass (Exponential class/stable)
6	Mat	Mat (Matern)
7	Ste	Mat (Matern M. Stein's parameterization)
8	Cir	Cir (circular)
9	Lin	Lin (linear)
10	Bes	Bes (bessel)
11	Pen	Pen (pentaspherical)
12	Per	Per (periodic)
13	Wav	Wav (wave)
14	Hol	Hol (hole)
15	Log	Log (logarithmic)
16	Pow	Pow (power)
17	Spl	Spl (spline)
18	Leg	Leg (Legendre)
19	Err	Err (Measurement error)
20	Int	Int (Intercept)

Tabela B.5: Modelos permissíveis de variograma para o objeto vgm

Para utilizarmos a função vgm() primeiramente adicionamos como argumento inicial a contribuição da estrutura, o tipo de modelo (Esférico, Exponencial, etc), o alcance da estrutura e o efeito pepita. Podemos adicionar o parâmetro anis, ao qual contém primeiramente o azimute (em graus) da direção principal e o fator de redução do alcance para a elipse. Em outras palavras se o alcance máximo na direção principal é 50m, ao adicionarmos um fator de 0.6 fazemos com que a direção de menor continuidade seja de 30m. Para adicionarmos mais de uma estrutura podemos concatená-las com o argumento add.to, adicionando quantas estruturas forem necessárias para formar o modelo de continuidade espacial. Finalmente podemos plotar o gráfico utilizando o comando plot(). O código fonte abaixo demonstra como obter um modelo de variograma a partir dos dados do Walker Lake.

```
1
2
3   #Variancia da variavel V
4   var(walker$V)
5
6   #Variogramas experimentais
7
8   v.dir = variogram(V~1, walker, width=10, cutoff= 200, tol.hor=45, alpha =
        (0:7)*22.5 )
9
10  # Modelagem da primeira estrutura
11
12  v.anis1 = vgm(59929,"Sph",40,20000, anis=c(157,0.6))
13
```

```
14  # Modelagem da segunda estrutura
15
16  v.anis2 = vgm(20000,"Exp",100, 0, anis=c(157,0.6),add.to =v.anis1 )
17
18  # Plotagem do grafico
19  plot(v.dir, v.anis2)
```

Listing B.24: Criação de um vetor em R

A figura abaixo demonstra o modelo de continuidade espacial para o Walker Lake para a variável V. Neste caso a direção de maior continuidade foi considerada em 157.5 graus e duas estruturas foram adicionadas.

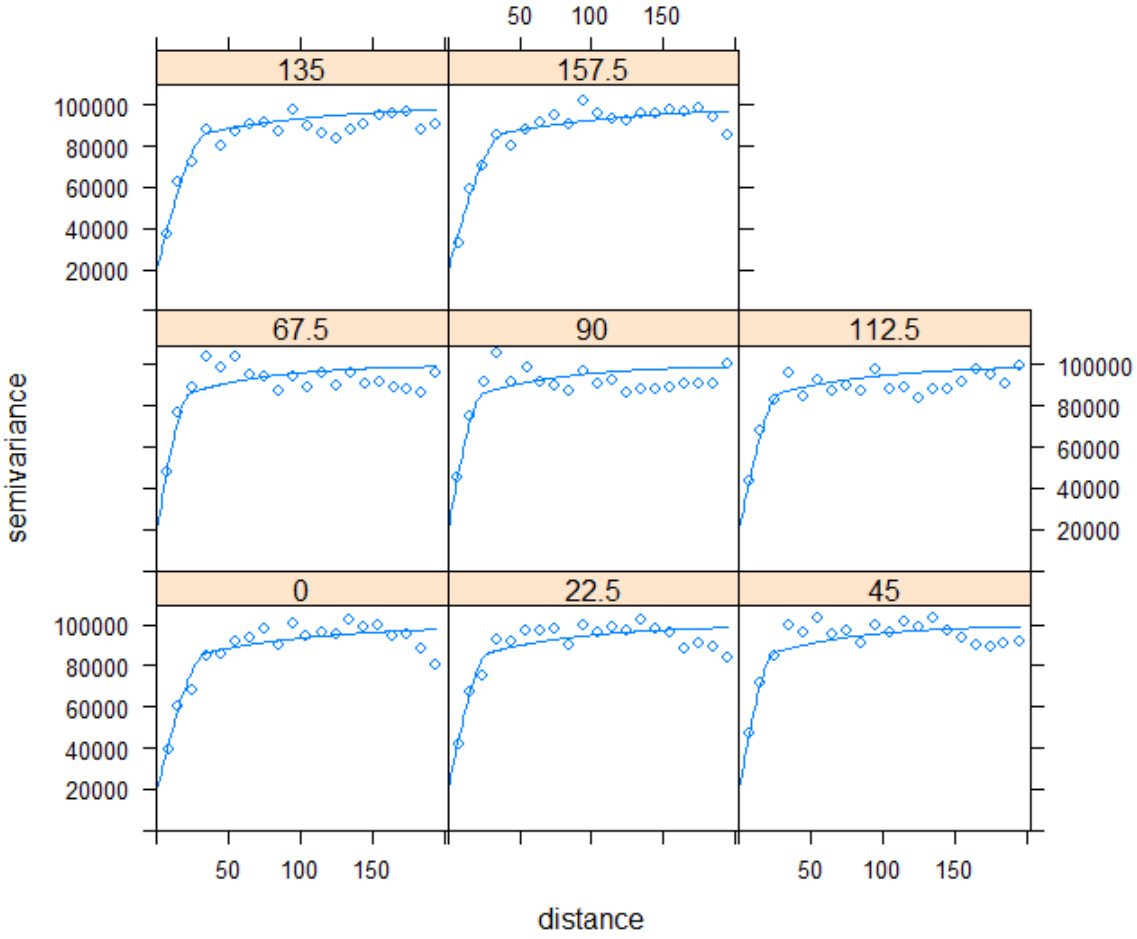


Figura B.12: Ajuste do modelo para diferentes direções utilizando a função vgm

B.22 Validação Cruzada

Para testar a eficiência de diferentes modelos de variograma, tal como diferentes estratégias de busca da krigagem, podemos utilizar a validação cruzada. De acordo com os erros exibidos pela validação, podemos modificar os parâmetros de krigagem e do variograma para obter os menores erros possíveis. O resíduo é uma medida adequada neste caso para o erro de estimativa, podemos plotar os resultados em um gráfico de bolhas. A função krige.cv() é calculada fornecendo primeiramente

a variável de interesse, o dataframe que está contido a variável, o modelo de variograma ajustado na seção anterior, o número mínimo de amostras utilizadas na krigagem, o número máximo de amostras utilizadas na krigagem, a máxima distância de procura dos dados e o número de dados retirados durante a validação para se computar o erro médio dos valores reais e krigados. O código fonte abaixo demonstra a validação cruzada.

```
1
2  # Validacao cruzada
3  cv = krige.cv(V~1, walker, v.anis2 , nmin= 3, nmax=10, maxdist=100, nfold=20)
4
5  # Sumario estatistico da validacao
6  summary(cv)
7
8  #Plotagem dos residuos da validacao cruzada
9  bubble(cv["residual"])
```

Listing B.25: Criação de um vetor em R

A figura abaixo demonstra os erros residuais a partir da validação cruzada do depósito Walker Lake.

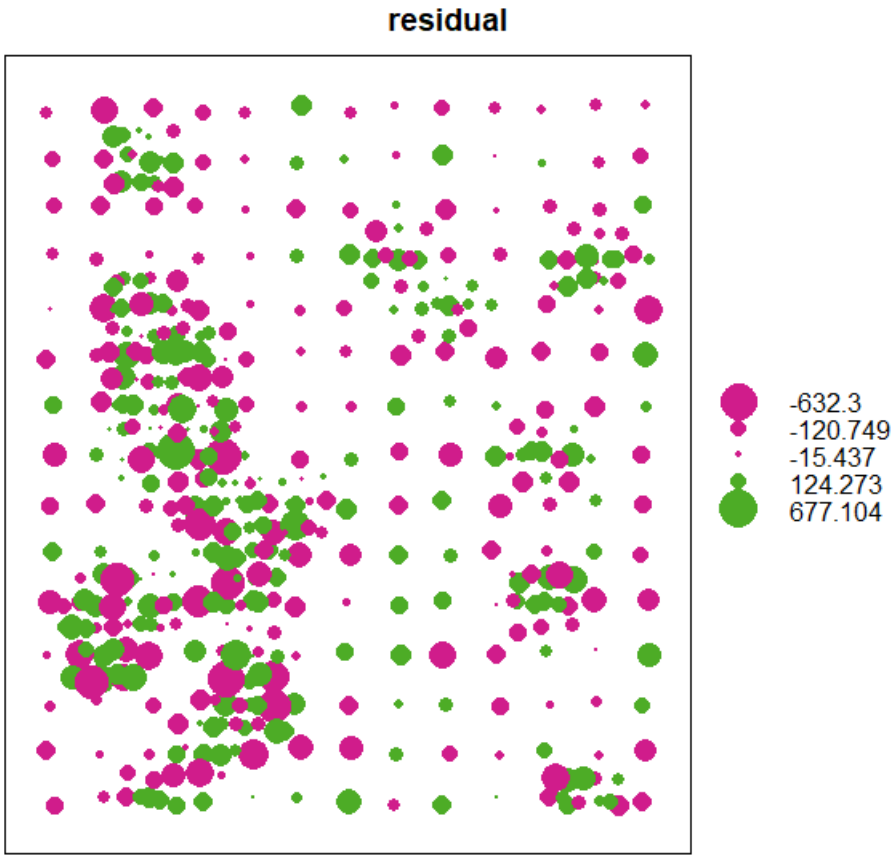


Figura B.13: Erros da validação cruzada demonstrados em um gráfico de bolhas

B.23 Krigagem

Encontrados os melhores modelos de variograma e estratégia de busca possíveis, podemos realizar a krigagem da variável de interesse V. Para isso criamos um grid assim como no vizinho mais próximo

utilizando os comandos já conhecidos. Então utilizamos o comando krige(), cujos argumentos são, primeiramente a variável de interesse a ser krigada, em seguida o dataframe em que esta variável está contida, o grid criado e os parâmetros da estratégia de busca, tais como mínimo número de amostras (nmin), máximo número de amostras (nmax), a máxima distância de procura (nmax) e finalmente o modelo de variograma ajustado. O código fonte abaixo demonstra a krigagem dos valores da variável V, do depósito do Walker Lake.

```
1
2  # Criar um grid
3  grid_stat = makegrid(walker, cellsize = 5)
4  grid_stat = SpatialPixels(SpatialPoints(grid_stat))
5
6  # Krigar os valores
7  kriged = krige(V~1, walker, grid_stat,nmin=2, nmax=3, maxdist=100, v.anis2)
8
9  # Plotar a variavel estimada
10 spplot(kriged['var1.pred'], scales = list(draw =T))
11
12 # Plotar a variancia de krigagem
13 spplot(kriged['var1.var'] , scales = list(draw =T))
14 summary(kriged)
```

Listing B.26: Criação de um vetor em R

O gráfico abaixo demonstra o valor krigado da variável V a partir da estratégia de busca e do variograma ajustado.

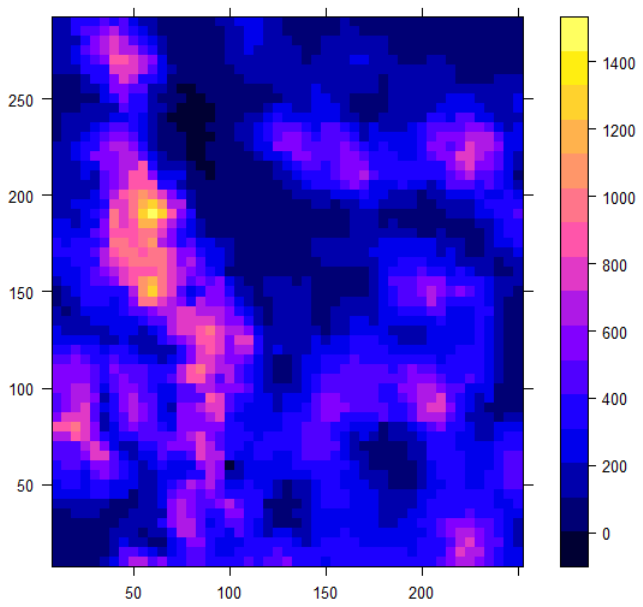


Figura B.14: Krigagem da variável V do depósito Walker Lake

O gráfico abaixo demonstra a variância de krigagem da variável V a partir da estratégia de busca e do variograma ajustado.

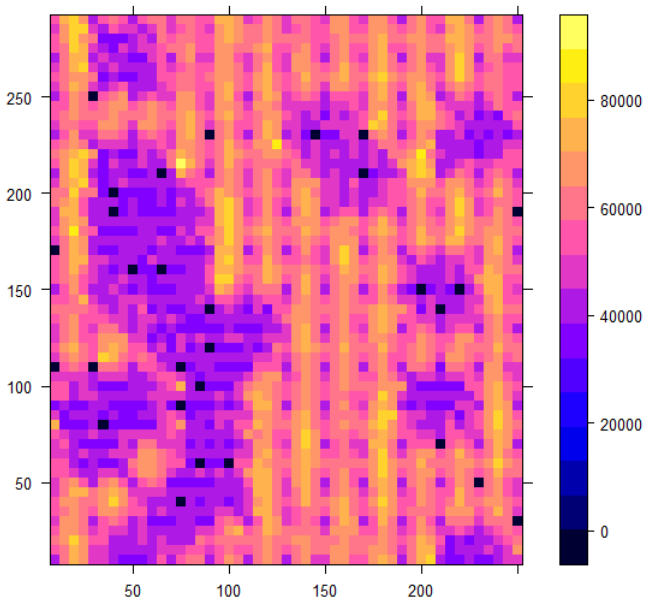


Figura B.15: Krigagem da variável V do depósito Walker Lake

Bibliography

Articles

[3] Noel Cressie. “Fitting variogram models by weighted least squares”. Em: *Journal of the International Association for Mathematical Geology* 17.5 (1985), páginas 563–586.

[7] Emmanuel Gringarten e Clayton V Deutsch. “Teacher’s aide variogram interpretation and modeling”. Em: *Mathematical Geology* 33.4 (2001), páginas 507–534.

[8] Edward H Isaaks e R Mohan Srivastava. “Applied geostatistics”. Em: (1989) (ver páginas 21, 118).

[10] Georges Matheron. “Principles of geostatistics”. Em: *Economic geology* 58.8 (1963), páginas 1246–1266 (ver página 21).

[11] AB McBratney, R Webster e TM Burgess. “The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I: Theory and method”. Em: *Computers & Geosciences* 7.4 (1981), páginas 331–334.

Books

[1] Isobel Clark. *Practical geostatistics*. Volume 3. Applied Science Publishers London, 1979.

[2] Timothy C Coburn, Jeffrey M Yarus, Richard L Chambers et al. *Stochastic modeling and geostatistics: principles, methods, and case studies, vol. II, AAPG computer applications in geology 5*. Volume 5. AAPG, 2005.

[5] Norman Richard Draper, Harry Smith e Elizabeth Pownell. *Applied regression analysis*. Volume 3. Wiley New York, 1966.

[6] Pierre Goovaerts. *Geostatistics for natural resources evaluation*. Oxford University Press on Demand, 1997 (ver página 21).

[9] Andre G Journel e Ch J Huijbregts. *Mining geostatistics*. Academic press, 1978.

[12] Alastair J Sinclair e Garston H Blackwell. *Applied mineral inventory estimation*. Cambridge University Press, 2002.

[13] Hans Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2013.