

LIÇÕES FUNDAMENTAIS DE  
GEOESTATÍSTICA

# Introdução a **Geoestatística**

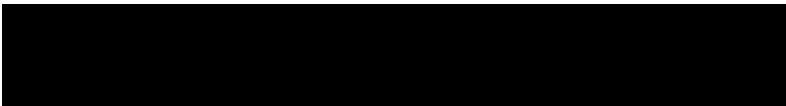
Com aplicações em R, GSLIB e SGEMS

**David A. Drumond**

**Fernanda G. F. Niagini**

**João Felipe C.L. Costa**

**Roberto M. Rolo**



# **Geoestatística**

**Introdução aos princípios e aplicações em R, GSLIB e SGeMS**

**David Alvarenga Drumond**

**Fernanda Gontijo Fernandes Niquini**

**Roberto Mentzingen Rolo**

**João Felipe Coimbra Leite Costa**

David Alvarenga Drumond  
Fernanda Gontijo Fernandes Niquini  
Roberto Mentzingen Rolo

# Geoestatística - Introdução aos princípios e aplicações em R, GSLIB e SGeMS

1 edição

Belo horizonte  
22/09/2017



*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*



# Conteúdo

<b>1</b>	<b>Prefácio</b>	<b>13</b>
<b>2</b>	<b>Introdução a geoestatística</b>	<b>17</b>
2.1	Introdução ao capítulo	17
2.2	Afinal, o que é geoestatística?	18
2.3	Qual é o objeto de estudo da geoestatística?	21
2.4	O que podemos fazer com a geoestatística?	23
2.5	Como utilizar a geoestatística?	26
2.6	O que a geoestatística não faz?	27
2.7	Questões éticas na avaliação de depósitos minerais	29
2.8	Alguns conceitos iniciais sobre jazidas minerais	30
2.8.1	Minério	30
2.8.2	Teor de corte e teor crítico	31
2.8.3	Continuidade	31
2.8.4	Diluição	32
2.8.5	Recursos e reservas minerais	32
2.8.6	Precisão e Exatidão	34
2.9	Conclusões	36
2.10	Exercícios	36

<b>3</b>	<b>Variáveis aleatórias regionalizadas .....</b>	<b>39</b>
3.1	Introdução ao capítulo	39
3.2	Variáveis aleatórias	40
3.3	Função de distribuição acumulada - fda	42
3.4	Função de densidade de probabilidade - fdp	44
3.5	Variáveis regionalizadas	44
3.6	Funções aleatórias	47
3.7	Hipótese de estacionaridade	52
3.8	Momentos estatísticos	55
3.9	Ergocidade	59
3.10	Homocedasticidade e heterocedasticidade	59
3.11	Relação Volume Variância	60
3.12	Conclusões	64
3.13	Exercícios	64
<b>4</b>	<b>Estatística univariada .....</b>	<b>67</b>
4.1	Introdução	67
4.2	Estatísticas pontuais	70
4.2.1	Medidas de tendência central .....	72
4.2.2	Medidas de posição .....	74
4.2.3	Medidas de dispersão .....	75
4.2.4	Assimetria .....	76
4.2.5	Coeficiente de variação .....	77
4.2.6	Conjugando estatísticas pontuais .....	78
4.3	Validação do banco de dados e valores outliers	79
4.4	Descrição espacial das amostras	83
4.5	Histograma	86
4.6	Inferência Estatística	90
4.6.1	Famílias de distribuições estatísticas .....	90
4.7	Distribuição t-Student	96
4.8	Dimensionamento de malhas regulares	97
4.9	Exercícios	98

<b>5</b>	<b>Estatística bivariada</b>	<b>99</b>
5.1	Introdução	99
5.2	Probabilidade condicional e Esperança condicional	100
5.2.1	Probabilidades condicionais e conjuntas	100
5.2.2	Esperança condicional	102
5.3	Ferramentas gráficas	104
5.3.1	Gráfico Q-Q plot	104
5.3.2	Gráfico p-p plot	106
5.3.3	Gráfico de dispersão	108
5.4	Régressão linear	110
5.5	Intervalo de segurança para a régressão linear	113
5.6	Régressão linear múltipla	115
5.7	Coeficiente de correlação	116
5.8	Exercícios	119
<b>6</b>	<b>Métodos clássicos e desagrupamento</b>	<b>121</b>
6.1	Introdução	121
6.1.1	Princípio da mudança gradual	122
6.1.2	Princípio dos pontos mais próximos	123
6.1.3	Princípio da generalização	124
6.2	Composição	125
6.3	Composição em seções verticais	126
6.4	Determinação de volumes	128
6.5	Inverso do quadrado da distância - IQD	129
6.6	Tesselação de Delunay	130
6.7	Polígonos de Thiessen	131
6.8	Estatísticas desagrupadas	135
6.8.1	Polígonos de influência	137
6.8.2	Desagrupamento por células	138
<b>7</b>	<b>Continuidade Espacial</b>	<b>141</b>
7.1	Definição de continuidade espacial e variografia	141
7.2	Procura de pares de amostras no espaço	144

<b>7.3</b>	<b>Funções experimentais de continuidade espacial</b>	<b>147</b>
7.3.1	Efeito dos dados sobre os valores experimentais .....	147
7.3.2	Funções de continuidade espacial mais comuns .....	147
7.3.3	Outras funções experimentais .....	149
7.3.4	Parâmetros de busca .....	152
<b>7.4</b>	<b>Modelagem de funções de continuidade espacial</b>	<b>154</b>
7.4.1	Modelos de variogramas permissíveis .....	154
7.4.2	Parâmetros das funções de continuidade .....	154
7.4.3	Modelos de continuidade espacial mais comuns .....	155
7.4.4	Anisotropia .....	157
7.4.5	Funções de continuidade espacial cruzadas .....	159
7.4.6	Modelo linear de correacionalização .....	160
7.4.7	Modelagem automática de variogramas .....	160
<b>8</b>	<b>Krigagem</b> .....	<b>163</b>
<b>8.1</b>	<b>Introdução</b>	<b>163</b>
<b>8.2</b>	<b>Krigagem Ordinária</b>	<b>166</b>
<b>8.3</b>	<b>Krigagem Simples</b>	<b>167</b>
<b>8.4</b>	<b>Krigagem de blocos</b>	<b>168</b>
<b>8.5</b>	<b>Influência nos pesos da krigagem</b>	<b>170</b>
8.5.1	Influência do modelo de continuidade espacial nos pesos .....	171
8.5.2	Influência dos parâmetros do variograma .....	171
8.5.3	Efeito da geometria das amostras .....	174
<b>8.6</b>	<b>Estratégia de procura</b>	<b>176</b>
<b>8.7</b>	<b>Validação da krigagem</b>	<b>178</b>
8.7.1	Verificação do comportamento dos mapas krigado e das amostras .....	178
8.7.2	Comparação da média global com a média das amostras .....	179
8.7.3	Análise de deriva de bandas do mapa .....	179
8.7.4	Validação cruzada .....	180
8.7.5	Verificação de pesos negativos .....	180
<b>9</b>	<b>Mudança de suporte</b> .....	<b>183</b>
<b>9.1</b>	<b>Mudança de suporte</b>	<b>183</b>
9.1.1	Correção afim .....	184
9.1.2	Transformação lognormal indireta .....	185

<b>9.2</b>	<b>Curva de teor e tonelagem</b>	<b>185</b>
9.2.1	Curvas de teor e tonelagem derivadas de histogramas das amostras .....	186
9.2.2	Curvas de teor e tonelagem a partir de distribuição de probabilidades contínuas das amostras .....	187
9.2.3	Curvas de teor e tonelagem baseadas na dispersão dos blocos estimados .....	187
9.2.4	Curvas de teor e tonelagem baseadas na estimativa dos blocos .....	188
9.2.5	Erros associados à determinação da curva de teor-tonelagem .....	188
<b>10</b>	<b>Estimativa x Realidade .....</b>	<b>189</b>
<b>10.1</b>	<b>Introdução</b>	<b>189</b>
10.1.1	Controle de teores do minério .....	190
10.1.2	Uso de fatores de comparação - forma clássica .....	190
10.1.3	Uso de fatores de comparação - forma probabilística .....	191
10.1.4	Críticas à geoestatística .....	192
<b>A</b>	<b>Geoestatística multivariada .....</b>	<b>195</b>
<b>A.1</b>	<b>Modelos multivariados</b>	<b>197</b>
A.1.1	Krigagem simples com médias locais variáveis .....	198
A.1.2	Krigagem com deriva externa .....	198
A.1.3	Cokrigagem .....	200
A.1.4	Influência dos dados secundários .....	202
A.1.5	Condição não tradicional e tradicional da cokrigagem .....	202
A.1.6	Cokrigagem Colocada .....	202
<b>B</b>	<b>Geoestatística utilizando o software R .....</b>	<b>203</b>
<b>B.1</b>	<b>Introdução</b>	<b>203</b>
<b>B.2</b>	<b>Instalação do R</b>	<b>205</b>
<b>B.3</b>	<b>RStudio</b>	<b>206</b>
<b>B.4</b>	<b>Noções preliminares</b>	<b>207</b>
<b>B.5</b>	<b>O R como uma calculadora</b>	<b>208</b>
<b>B.6</b>	<b>Utilizando funções no R</b>	<b>208</b>
<b>B.7</b>	<b>Operadores Relacionais</b>	<b>209</b>
<b>B.8</b>	<b>Operadores Lógicos no R</b>	<b>209</b>
<b>B.9</b>	<b>Pedindo ajuda no R</b>	<b>210</b>
<b>B.10</b>	<b>Pacotes do R</b>	<b>211</b>

<b>B.11</b>	<b>Criando vetores</b>	211
<b>B.12</b>	<b>Condicional</b>	213
<b>B.13</b>	<b>Repetições</b>	213
<b>B.14</b>	<b>Concatenação de funções</b>	214
<b>B.15</b>	<b>DataFrames</b>	214
<b>B.16</b>	<b>Mapa de localização</b>	215
<b>B.17</b>	<b>Histogramas</b>	218
<b>B.18</b>	<b>Boxplots</b>	219
<b>B.19</b>	<b>Regressão Linear</b>	221
<b>B.20</b>	<b>Vizinho mais próximo</b>	223
<b>B.21</b>	<b>Variograma</b>	225
<b>B.22</b>	<b>Validação Cruzada</b>	231
<b>B.23</b>	<b>Krigagem</b>	233
<b>C</b>	<b>Geoestatística utilizando o GSLib</b>	237
<b>C.1</b>	<b>Introdução</b>	237
<b>C.2</b>	<b>A execução do GSLIB</b>	238
<b>C.3</b>	<b>Entrada de dados</b>	238
<b>C.4</b>	<b>Exemplos de aplicação do GSLIB</b>	239
<b>C.4.1</b>	<b>Criando um histograma com o HISTPLT</b>	239
<b>C.4.2</b>	<b>Criando um gráfico de dispersão com o SCATPLT</b>	242
<b>C.4.3</b>	<b>Criando um mapa de localização com o LOCMAP</b>	243
<b>C.5</b>	<b>desagrupamento utilizando células móveis com o DECLUS</b>	245
<b>C.6</b>	<b>Convenção da orientação de eixos de anisotropia do GSLIB</b>	247
<b>C.7</b>	<b>Variograma experimental (GAMV/ VARGPLT)</b>	248
<b>C.8</b>	<b>Modelagem de variogramas (VMODEL/VARGPLT)</b>	251
<b>C.9</b>	<b>Validação Cruzada com (KT3D/LOCMAP)</b>	254
<b>C.10</b>	<b>Krigagem com (KT3D/PIXELPLT)</b>	258
<b>D</b>	<b>Geoestatística utilizando o SGeMS</b>	263
<b>D.1</b>	<b>Importando um arquivo de pontos no SGeMS</b>	264
<b>D.2</b>	<b>Visualização dos dados - Mapa de localização</b>	267
<b>D.3</b>	<b>Criação do histograma</b>	268

**Bibliografia .....** 271





## 1. Prefácio

*Se seus problemas têm solução,  
aprendes a solucioná-los, se não têm  
solução, aprendes a não preocupar.  
Os problemas só surgem quando não  
se aprende a agir, no primeiro caso  
aja, no segundo contemple.*

...

A geoestatística é, sem dúvida, uma das mais belas ferramentas para trabalharmos com incerteza em projetos que envolvem análise espacial e engenharia. Não é somente um ponto de partida para analisar e avaliar, mas uma proposição da humildade humana, nossa capacidade da incompreensão extensiva do universo a nossa volta.

Quando pensamos na ciência tradicional, desenvolvida nos primórdios do século XVIII com o advento do iluminismo, vemos claramente o ser humano tentando desenvolver ferramentas para explicar o universo em suas nuâncias, criando modelos físicos e matemáticos das representações de problemas reais. Estes modelos, sólidos e claros, possuem uma capacidade incrível de reproduzibilidade, podendo ser aplicados em diferentes contextos e situações. Apesar desta maleabilidade, em certos momentos problemas complexos nos foram apresentados, sendo incapazes de serem descritos pelas formas simples por estes antigos modelos.

As técnicas geoestatísticas, como as demais ferramentas modernas de estatística, marcam um ponto na história, quando assumimos a impossibilidade de entender todos os processos, mas que possamos descrevê-los de forma verossímil, criando modelos matemáticos que se aproximam da realidade, desconhecida e muitas vezes intangível. Os modelos estatísticos neste caso não são reproduutíveis, sendo impossível a um avaliador aplicar o mesmo modelo para diferentes depósitos minerais. No entanto, as técnicas são simples e eficientes, reproduzindo padrões sobre a geologia e fenômenos espaciais de forma eficiente.

O surgimento da geoestatística está diretamente relacionado aos trabalhos do professor [George Matheron](#) a partir de análises sobre estudos estatísticos de [Singel](#) em minas de ouro da África do Sul. Depósitos de ouro são conhecidos pela sua complexidade geológica, apresentando alta variabilidade em suas propriedades químicas e geológicas. Os métodos clássicos de avaliação de recursos se demonstraram inefficientes para lidar com esta complexidade do problema, pois consideravam apenas questões geométricas de disposição das amostras, sem considerar sua interdependência espacial. Surgiu então a geoestatística, que pretendia adicionar informação não apenas pelo posicionamento de amostras, mas pela sua dependência espacial.

Diferentemente dos métodos estatísticos clássicos, em que consideramos as amostras independentes entre si, na geoestatística consideramos que o posicionamento espacial, volume das amostras e orientação possuem forte relação com as avaliações que realizamos. Esta proposição torna os métodos geoestatísticos ainda eficientes por mais de 50 anos de desenvolvimento nas avaliações de recursos minerais, pois corrobora com a questão física de formação dos controles geológicos destes depósitos.

Após anos de desenvolvimento nas técnicas geoestatísticas, diferentes pesquisas e conhecimentos foram derivados das técnicas iniciais de Matheron. Mesmo assim, esta disciplina ainda parece obscura nos cursos de engenharia de minas, geologia e geografia não somente no Brasil como no mundo. Quando focamos no contexto brasileiro, o problema se acentua, pois são poucas as bibliografias escritas em português sobre este assunto. Pensando nisso, nós alunos do curso de pós-graduação Engenharia de Minas da Universidade Federal do Rio Grande do Sul, juntamente com a orientação dos profissionais da universidade, iniciamos este livro como um projeto para uma série de publicações sobre [Geoestatística](#), [Planejamento de Lavra](#), [Amostragem de jazidas](#), [Métodos de cubagem](#) entre outras disciplinas da mineração.

A abordagem adotada neste livro de [Introdução a Geoestatística](#) envolve a chamada geoestatística linear clássica, que envolve desde métodos tradicionais de geoestatística, de caracterização da continuidade espacial e da krigagem, principalmente ordinária e simples, que envolvem principalmente os primeiros trabalhos desenvol-

vidos na década de 70 pelo professor George Matheron. Ao final do livro temos seções dedicadas exclusivamente para a apresentação da geoestatística em softwares gratuitos e linguagem de programação em R. O livro se desenvolve nos seguintes capítulos:

- **Capítulo 2:** Este capítulo apresenta uma introdução a respeito da ciência da geoestatística, explicando os principais conceitos para entendermos algumas questões chaves, como *O que é a geoestatística?*, *O que podemos fazer com a geoestatística?*, *Como podemos utilizar a geoestatística?*
- **Capítulo 3:** Apresenta o objeto principal do estudo da geoestatística, a teoria das variáveis regionalizadas. Apresentamos a conceituação e o formalismo matemático deste conjunto de técnicas.
- **Capítulo 4:** Inserimos conceitos iniciais sobre estatística univariada, ou seja, as técnicas utilizadas para avaliação de apenas uma única variável. Para isso mostramos estatísticas gráficas, estatísticas pontuais, enfocando principalmente no contexto da mineração.
- **Capítulo 5:** Inserimos conceitos iniciais sobre estatística multivariada, ou seja, aquela que analisa informações conjuntas de duas ou mais variáveis. São apresentadas estatísticas pontuais, estatísticas gráficas, entre outras ferramentas específicas deste assunto.
- **Capítulo 6:** Neste capítulo apresentamos as principais formas de análise de agrupamento para amostras espaciais. O objetivo destas técnicas é criar condições para reduzir o efeito de amostragens localizadas e ponderar as estatísticas para que indiquem sua representação espacial verdadeira.
- **Capítulo 7:** Introduzimos os conceitos de continuidade espacial e dependência espacial das amostras, mostrando por exemplo, a estimativa e modelagem de funções variograma e covariograma.
- **Capítulo 8:** Apresentamos as principais técnicas de estimativa geoestatística linear, como krigagem simples e ordinária.
- **Capítulo 9:** Introduzimos os conceitos de mudança de suporte, e as principais ferramentas para analisar a mudança de volumes estimados.
- **Capítulo 10:** Apresentamos as principais técnicas utilizadas para avaliar os modelos krigados e estimativas realizadas, como por exemplo, as curvas de teor e tonelagem comumente utilizadas nos trabalhos de avaliação de recursos.

A seção final de cada capítulo apresenta uma série de exercícios com arquivos de dados apresentados no material de apoio deste livro. Ao final destes capítulos apresentamos os apêndices A, B e C, com avaliações geoestatísticas realizadas no depósito *Walker Lake*. Os dados deste depósito podem ser encontrados junto com o material de apoio.



## 2. Introdução a geoestatística

*A estimativa de recursos é o processo de criação de um reflexo tridimensional da mineralização in situ baseado em amostras esparsas utilizando conhecimento geológico corrente e um caminhão carregado de senso comum.*

*The art and the science of resource estimation  
Jacqui Coombes*

### 2.1 Introdução ao capítulo

Este capítulo inicial pretende demonstrar os primeiros passos para entender a geoestatística. Afinal, o que é a geoestatística? Qual é o seu objeto de estudo? O que podemos fazer ou não com a geoestatística? Para que um marceneiro possa fazer uma cadeira, por exemplo, ele precisa entender de suas ferramentas e de seu funcionamento, para que possa selecionar as mais adequadas para o seu trabalho. Entendendo o que é a geoestatística e como podemos utilizá-la, principalmente no setor da mineração, estabelecemos um vínculo necessário para a aplicação correta desta poderosa ferramenta.

## 2.2 Afinal, o que é geoestatística?

A importância das substâncias metálicas na indústria mineral brasileira é historicamente associado aos tempos da Colônia, procurando rotas inicialmente no estado de Minas Gerais. Segundo o relatório da Agência Nacional de Mineração (2018), a produção das principais substâncias metálicas no país atingiram um valor de 41,7 bilhões de reais em exportação. A produção mineral rende impostos para as regiões produtoras, que ao mesmo tempo permitem o desenvolvimento da economia local e geração de empregos. A decisão da extração e produção mineral, no entanto, advém do conhecimento geológico e de estimativas dos corpos minerais aos quais muitas vezes não se possui acesso. Os corpos geológicos, muitas vezes, não apresentam informações superficiais de fácil acesso, como afloramentos que permitem a definição da **atitute** das camadas, por isso é necessário realizar amostragens em grande profundidade como na obtenção de **testemunhos de sondagem**. A partir de uma sonda são retirados fragmentos de rocha à profundidades de até 400m, permitindo que tenhamos informações da composição direta das rochas naquela região. A figura 2.1 exemplifica a forma cilíndrica apresentada pelo testemunho de sondagem obtido em campanhas de pesquisa mineral.

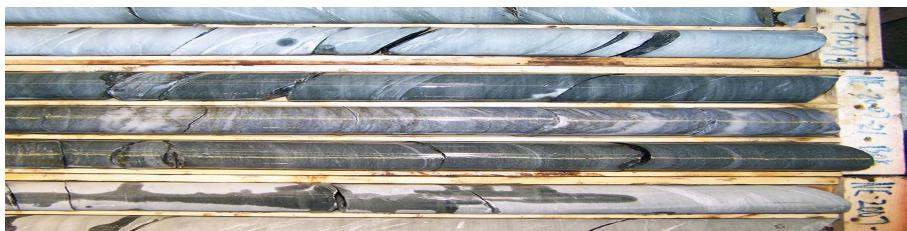


Figura 2.1: Testemunhos de sondagem de rochas. Amostras retiradas a partir da perfuração do solo, que apresentam uma informação contínua vertical das rochas e mineralogias presentes em uma região.

Desta forma, a única informação que possuímos é a informação vertical fornecida pelos testemunhos, como demonstrado na figura 2.2. As decisões da mineração não podem ser estabelecidas sem o conhecimento das informações entre os furos de sondagem, que podem representar malhas espaçadas em muitos metros. A amostragem exaustiva dos depósitos minerais também é inviável economicamente, pois em alguns casos, cada metro de amostra sondada pode corresponder a um valor de \$100 a \$400 reais.

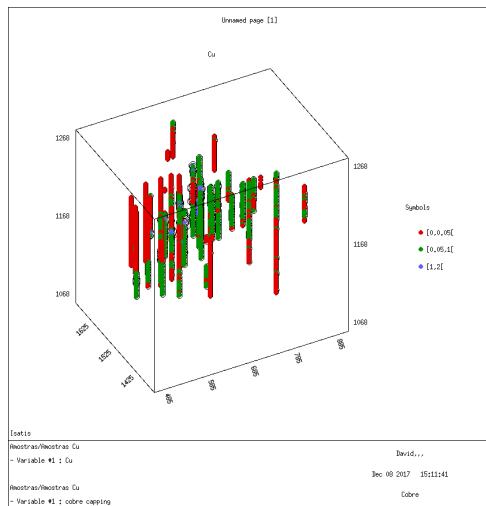


Figura 2.2: Testemunhos de sondagem representados em um software de mineração, para um depósito de cobre.

A geoestatística é a ciência que permite a espacialização das informações obtidas em um volumes menores, para um domínio maior, de forma a permitir o planejamento e tomada de decisões na mineração, e o estudo sistemático dos corpos mineralizados. A figura 2.3 demonstra a espacialização dos dados obtidos na figura 2.2 dos testemunhos de sondagem de cobre.

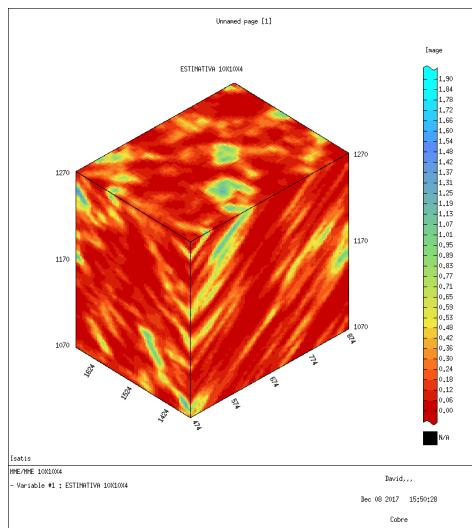


Figura 2.3: Espacialização a partir das amostras do testemunho de sondagem obtidos na figura 2.2, para um depósito de cobre.

A partir de um modelo espacializado é possível planejar a mineração e tomar a decisões na lavra, como a criação de **modelos econômicos**, a determinação das **cavas matemáticas**, o **sequenciamento das operações**. Segundo Rossi and Deutsch [2013], o objetivo principal da geoestatística consiste em 4 etapas principais:

1. Obtenção de amostras e administração da amostragem
2. Interpretação geológica e modelagem
3. Interpolação dos teores
4. Acesso às incertezas geológicas

A **obtenção de amostras e administração da amostragem** consiste no conjunto de técnicas utilizada para obter uma malha de amostragem que reduza o erro obtido pela interpolação espacial das propriedades de interesse. Por exemplo, a malha de amostragem em um depósito de ferro bandado, conhecido como (*Banded Iron Formations*) , pode ser dimensionada para que se reconheçam os minerais deletérios durante a fase de metalurgia.

A **interpretação geológica e modelagem** permite reconhecer as dimensões e formas do corpo geológico e principais estruturas. Um dos grandes desafios da mineração é conseguir reconhecer os limites dos corpos minerais e sua forma. A geoestatística permite utilizar técnicas que auxiliem no reconhecimento das formas destes corpos de maneira grosseira, a partir da *modelagem implícita*. O objetivo do uso destas técnicas é reduzir a quantidade de trabalho demandada pelo geólogo para que se possam produzir modelos condizentes com a realidade, e ao mesmo tempo, poupar trabalho excessivo pelo desenvolvimento de seções verticais.

A **Interpolação de teores** consiste no objetivo principal deste livro, em que realizamos a espacialização das propriedades de interesse das amostras para um domínio espacial maior. Esta espacialização pode ser realizada para apenas uma variável (caso univariado), ou para diversas variáveis em conjunto (caso multivariado). O objetivo principal da interpolação é garantir, com maior segurança possível, que um volume direcionado da lavra para o beneficiamento mineral possua **valor esperado**, ou **valor médio**, correto.

O **Acesso às incertezas geológicas** pode ser realizado a partir de técnicas avançadas de geoestatística como a simulação, ou geoestatística não-linear. Pretende-se desta forma tentar reconhecer as incertezas locais de uma propriedade do depósito, e avaliar quão díspares podem ser as medições em regiões do depósito que desconhecemos. A incerteza geológica é, sem dúvida, um dos fatores que mais afetam o **risco** do empreendimento mineiro. Para que os investidores possam verificar o risco de seus investimentos, foram criados os **códigos de mineração**, que criaram padrões nomear regiões do depósito mineral com maior ou menor incerteza quanto uma propriedade de interesse, geralmente aquela de retorno econômico.

Segundo [Matheron \[1963\]](#), criador da geoestatística, podemos definí-la tal como:

**R** "Geoestatística, na sua maior aceitação, consiste no estudo da distribuição do espaço de valores úteis para engenheiros de minas e geólogos, como teores, espessura da camada, ou acumulação, incluindo as práticas mais importantes para a avaliação de depósitos minerais- [Matheron \[1963\]](#)

Atualmente o uso da geoestatística compreende uma diversidade enorme de áreas, desde a **engenharia civil**, **engenharia agrícola**, **engenharia ambiental**, **geografia**, **engenharia hídrica** e até mesmo em áreas que não se resumem à dados geograficamente referenciados, mas espacialmente referenciados em objetos ou seres, como a **mecânica** ou **medicina**. Podemos entender a geoestatística sob uma perspectiva mais ampla, abordando o estudo das incertezas a cerca de fenômenos temporalmente ou espacialmente localizados. O professor [Goovaerts \[1997\]](#), demonstra claramente a nossa dificuldade de entender as incertezas:

**R** "A respeito da incerteza ... ela surge do nosso conhecimento imperfeito do fenômeno, dependente dos dados e ainda mais dependente do modelo, em que o modelo especifica nossas decisões (concepções) *a priori* do fenômeno. Nenhum modelo tal como a medida da incerteza, pode ser objetiva.- [Goovaerts \[1997\]](#)

Podemos entender então as limitações acerca dos modelos geoestatísticos. Estamos sempre **dependentes das amostras recolhidas para a avaliação**, como também a escolha dos modelos que melhor representam as características de um fenômeno. A geoestatística constitui atualmente a área que melhor consegue caracterizar a incerteza geológica, dada as condições de amostragem que obtemos na mineração e de muitos problemas georeferenciados. Definimos a geoestatística como:

**Definição 2.2.1 — Geoestatística.** *A geoestatística é a ciência capaz de transformar as informações obtidas por amostras georeferenciadas em conhecimento, a partir da caracterização da incerteza geológica, da interpretação destes dados, das inferências e estimativas, e da tomada de decisão pelo reconhecimento do fenômeno estudado.*

## 2.3 Qual é o objeto de estudo da geoestatística?

A geoestatística é a ciência que permite o estudo de variáveis regionalizadas. O capítulo 3 trará informações a respeito desta teoria, que compõe o objeto principal do estudo da geoestatística. Uma variável regionalizada é aquela que pode assumir um valor específico no espaço. Este valor é **determinístico**, gerado a partir de fenômenos que muitas vezes não conhecemos. Por não conseguirmos acessar as informações a respeito desta variável, optamos por utilizar uma metodologia **estocástica** para acessar a nossa **incerteza** a cerca do fenômeno que estudamos. Desta

forma pensamos na variável regionalizada com um aspecto **dicotômico**, a medida que possui valor real onde conhecemos, e valor aleatório onde desconhecemos.

O físico Erwin Schrödinger desenvolveu um problema em 1935 muito similar a esta condição das variáveis regionalizadas. O experimento foi chamado de "Gato de Schrödinger". O experimento propunha que um gato fosse preso em uma caixa, com um veneno que poderia aleatoriamente ser liberado, matando o gato em seguida. Para quem observa a experiência do lado de fora, não há como detectar se o gato está vivo ou morto, logo o estado de sobrevivência do gato é indefinido, dependendo da real observação de dentro da caixa. A figura 2.4 demonstra este experimento.

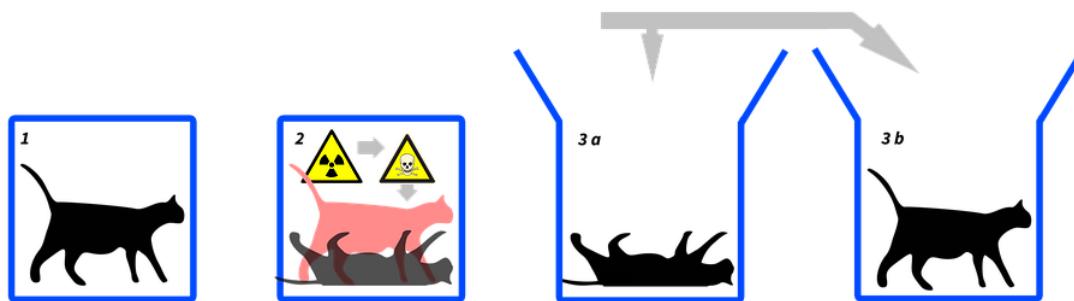


Figura 2.4: Experiência do "Gato de Schrödinger", proposta em 1935 pelo físico Austríaco Erwin Schrödinger. 1) O gato está dentro da caixa, 2) Passado um tempo o observador externo não sabe o que está dentro da caixa. 3a) Abre-se a caixa e define-se que o gato está morto. 3) Abre-se a caixa e define-se que o gato está vivo.

Quando pensamos em termos da mineração, o problema de se definir minério ou estéril é similar ao do "Gato de Schrödinger". Não sabemos de fato o que será minerado deve ser enviado para o beneficiamento mineral, a não ser que de fato retirarmos o material do local. As **variáveis regionalizadas** funcionam de forma bem semelhante, possuindo este aspecto ao mesmo tempo determinístico e aleatório. Um depósito mineral é um evento geológico realizado durante milhões de anos. Durante o tempo de existência humana é quase impossível que estes depósitos minerais se modifiquem. Desta forma o corpo mineral, ou o "gato" já está dentro da caixa há muito tempo, porém nos é impossível determinar o seu atual estado sem que ocorra a mineração.

Ao estudar as variáveis regionalizadas, a geoestatística propõe encontrar valores e relações que desconhecemos, sem obtermos informações diretas do depósito mineral. Considerando a variabilidade inerente destas variáveis podemos criar modelos estatísticos que possam inferir propriedades que desconhecemos em outras regiões do depósito mineral.

## 2.4 O que podemos fazer com a geoestatística?

A partir da avaliação dos depósitos minerais utilizando a geoestatística, podemos caracterizar o fenômeno espacial e quantificar as incertezas para diferentes **variáveis**. Uma variável é uma característica de interesse de estudo no depósito mineral, que se modifica segundo seu posicionamento no espaço. No estudo da mineração possuímos uma série de diferentes tipos de variáveis associadas ao depósito mineral, tais como:

1. **Químicas:** Teores de elementos químicos de interesse, ou de elementos deletérios prejudiciais no processamento mineral
2. **Físicas:** Dureza, densidade, condutibilidade térmica, condutibilidade hidráulica, saturação
3. **Geológicas:** Litologia, composição mineralógica, número de falhas, RQD (Rock quality index)
4. **Processamento:** Recuperação metalúrgica, recuperação mássica, moabilidade, consumo de reagentes
5. **Operacionais:** Resistência a penetração, consumo de explosivos, tempo de carregamento
6. **Econômicas:** Preço de mercado, valor presente líquido

O entendimento de cada uma destas variáveis permite a tomada de decisão de lavra de uma parte constituinte do depósito mineral. O uso de modelos geoestatísticos para cada uma destas variáveis deve, no entanto, ser específico para cada tipo de variável calculada. Neste livro introdutório abordamos principalmente os problemas que se relacionam com variáveis consideradas **aditivas**. Algumas variáveis que podem ser consideradas aditivas, e que representam o maior escopo de trabalho dos avaliadores de depósito mineral são **teor do elemento metálico, quantidade de metal de interesse, massa do minério e acumulação**. [Carrasco et al. \[2008\]](#) demonstra o conceito de aditividade de variáveis, expresso por:

 "Quantidades consideradas aditivas são aquelas que a quantidade média é igual a média das quantidades.- [Carrasco et al. \[2008\]](#)"

A variável teor, por exemplo, pode ser considerada uma variável aditiva, pois a média aritmética dos teores de duas regiões de mesma forma e volume é idêntico ao valor médio do teor nestas regiões. No caso da recuperação metalúrgica, por

exemplo, não há possibilidade de se considerar a média aritmética como valor médio, sendo impossível utilizar a geoestatística linear para estimar valores de recuperação metalúrgica. Carrasco et al. [2008] ainda afirma a necessidade de variáveis aditivas para se realizar o processo de estimativa diretamente por meio da geoestatística linear clássica.

**R** "Uma quantidade dita não aditiva, não pode modelar sua variabilidade espacial ou estimar diretamente- Carrasco et al. [2008]

Para estimar ou avaliar variáveis ditas não-aditivas recorremos aos métodos de **geoestatística não linear** ou **simulação geoestatística**. Estes métodos não serão abordados neste volume deste livro, apenas abordaremos os conceitos primários da *geoestatística linear*, que envolvem o tratamento de variáveis aditivas. No entanto, para o leitor iniciante, é importante entender que os métodos deste livro apenas se aplicam para **variáveis aditivas** e para amostra com o mesmo volume e forma, conceito denominado de **suporte amostral**. As estimativas realizadas no depósito mineral geralmente são feitas em volumes maiores, chamado de **suporte da estimativa** e compõe a chamada **unidade seletiva de lavra**. Segundo Rossi and Deutsch [2013] uma unidade seletiva de lavra pode ser caracterizada como:

**R** "Mínimo volume de material ao qual o minério e o estéril podem ser separados, em função do método de lavra e da seletividade- Rossi and Deutsch [2013]

Podemos então discretizar o espaço em pequenos blocos, para se realizar a estimativa nestes locais. Este é chamado de **modelo de blocos**, representado na figura 2.5.

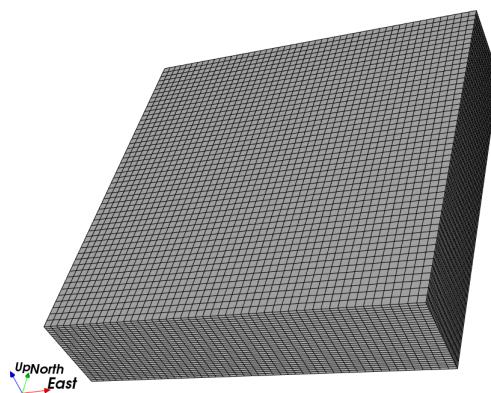


Figura 2.5: Representação de um modelo de blocos e discretização do espaço.

Assim podemos dividir o espaço a ser estimado em pequenos volumes de decisão na lavra. Para fins de planejamento mineral, quanto menor o tamanho destes blocos,

melhor a facilidade do planejamento. No entanto, para fins de avaliação de depósitos, blocos de tamanho pequeno produzem estimativas espúrias.

O equilíbrio entre estas duas vontades deve ser encontrado para constituir o tamanho adequado da unidade seletiva de lavra. Uma das regras de ouro da mineração geralmente afirma que: **o tamanho do bloco não deve ser inferior a 1/4 do tamanho da malha de amostragem**. Esta é uma afirmação atrela o tamanho do bloco geralmente a uma malha de amostragem bem definida e calculada, o que muitas vezes não condiz com as questões práticas.

**Proposição 2.4.1** Segundo a regra de ouro da geoestatística um bloco estimado não deve ter tamanho inferior a 1/4 do espaçamento da malha de amostragem. Quando considerada uma malha irregular este tamanho não pode ser menor que 1/4 do valor esperado dos espaçamentos. O valor esperado pode ser calculado a partir da média aritmética dos espaçamentos.

A discretização do depósito mineral em diferentes domínios nem sempre ocorre somente em modelos de blocos. Diferentes formas de caracterização dos volumes no espaço pode ser utilizada na geoestatística e no planejamento minera. A Figura (2.6) demonstra alguns exemplos de divisão do espaço. Algumas delas como **polígonos de influência** e **triangulação de Delunay** representam antigas formas de estimativa de um depósito mineral, mas que, no entanto, ainda são usuais por outras formas de análise.

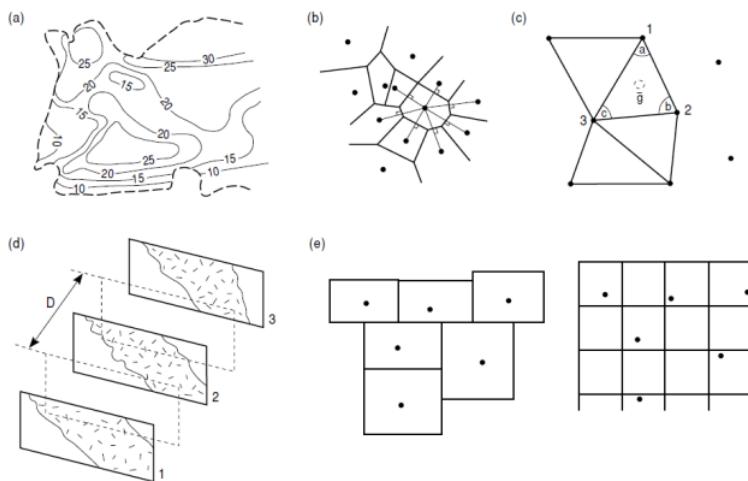


Figura 2.6: Figura demonstrando diversas apresentações de uma propriedade do depósito mineral. a) isolinhas b) Polígonos de influência c) triangulação d) seção paralelas e) blocos irregulares f) blocos regulares

A partir da definição das variáveis do depósito mineral o engenheiro é capaz de estimar a viabilidade técnica e econômica do depósito mineral. O artigo 6, do

decreto 9.406 de 12 de junho de 2018 do código mineral vigente define:

**Definição 2.4.1 — Jazida Mineral.** *Toda a massa de substância mineral ou fóssil, que aflore na superfície ou já exista no solo, no subsolo, no leito ou no subsolo do mar territorial, da zona econômica exclusiva ou da plataforma continental, que tenha valor econômico*

A partir da avaliação de depósitos minerais, conseguimos então, identificar regiões econômicas e capazes do aproveitamento industrial.

## 2.5 Como utilizar a geoestatística?

A geoestatística clássica utilizada neste livro não aborda conceitos matemáticos complexos, mas ainda sim constitui base para a resolução de muitos problemas de estimativa na mineração. Mesmo que os cálculos não sejam complexos, são de certa forma muito onerosos computacionalmente. As resoluções podem ser demoradas e alguns casos exigirem alta performance computacional. Desta forma, o uso de algoritmos refinados se torna cada vez mais importante nas análises geoestatísticas.

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 custosos, mas possuem suporte técnico e manutenção de seus sistemas. Apresentam código fechado ao público externo e pertencente geralmente aos 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](http://www.gslib.com) ou no guia de uso [DeutschCV \[1998\]](#). Neste livro abordamos nas seções A, B e C o uso dos principais softwares e linguagem R.

O alinhamento da geoestatística com o desenvolvimento de algoritmos cria dependências com as disciplinas de programação. Os engenheiros e geólogos estão cada vez mais alinhados com o desenvolvimento de algoritmos, principalmente com as linguagem R e Python, pela sua simplicidade e facilidade de implementação.

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 as implementações dos programas.

Entre as aplicações mais comuns encontradas em softwares de mineração, 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 quantidade de ferramentas adicionadas nestes programas geralmente os tornam pouco específicos para análises espaciais, obtendo apenas algumas rotinas específicas para trabalhos mais simples.

## 2.6 O que a geoestatística não faz?

Toda ferramenta possui suas limitações. A geoestatística é a melhor ferramenta para análises espaciais até então criada, mas ela possui limitações no uso de seus modelos. Primeiramente **a geoestatística não é uma caixa preta**. Isso significa que uma boa análise do depósito mineral não depende exclusivamente de apertar um botão, como muitos modelos mais simples fazem. Para criar modelos geoestatísticos adequados eles devem passar por uma intensa avaliação e reavaliação dos parâmetros de ajuste destes modelos.

Em segundo lugar **a geoestatística não é uma bola de cristal**. Quando realizamos estimativas estamos susceptíveis a erros e incertezas. Uma boa estimativa dos depósitos minerais propiciará redução dos erros, mas inevitavelmente não podemos sempre esperar resultados exatos. Além das condições relacionadas a escolha e refinamento dos modelos, também temos condições inerentes da incerteza geológica. [Maranhao \[1985\]](#) demonstra a classificação de jazidas minerais no ponto de vista da avaliação de reservas, identificando quatro grupos principais:

1. **Grupo 1.** Pertence aos depósitos estratiformes, cujos representantes típicos são as jazidas sedimentares de origem marinha, que possuem grandes dimensões, forma mais ou menos constante, e regularidade na distribuição de teores. Também incluem neste grupo as jazidas metamórficas de ferro, tais como nos depósitos do quadrilátero ferrífero. Também apresenta alguns depósitos de disposição horizontal ou a subhorizontal como jazidas de calcário, carvão, sais, gipsita e alguns depósitos para construção civil, como gnaisses e granitos.
2. **Grupo 2.** O segundo grupo apresenta corpos minerais interrompidos ou levemente interrompidos e uma distribuição mais irregular dos teores que do primeiro grupo. Estas representam as jazidas de alteração superficial como depósitos de níquel e bauxita, depósitos com pequenas intrusões alcalinas, como carbonatitos e sienitos, jazidas de rochas ultrabásicas e hidrotermais.
3. **Grupo 3** O terceiro grupo geralmente enquadra jazidas de forma variável e mineralização muito irregular, compondo os principais depósitos auríferos, platinoides e diamantes. Também aborda os depósitos de veios polimetálicos e depósitos de forma lenticular, como de cobre e níquel.
4. **Grupo 4** Representa o grupo mais irregular de todos, compondo pegmatitos de pedras preciosas, alguns veios hidrotermais com metais raros e nobres e algumas jazidas ultrabásicas de platina e diamante.

Quanto maior a irregularidade do depósito e heterogeneidade de suas propriedades, maior será a dificuldade dos modelos geoestatísticos de predizerem com exatidão os resultados. Dependendo da **continuidade espacial** da propriedade e da sua **dispersão**, a aplicação de modelos simples ou complexos simplesmente não altera o nosso conhecimento sobre a **incerteza geológica**, pois a complexidade do depósito mineral é tão grande, e as amostragens realizadas em tão pouca quantidade, que se torna mais fácil jogar uma moeda para cima para decidir se devemos ou não lavrar um depósito. Neste caso, quando os métodos geoestatísticos falham,

é necessário rever as metodologias de amostragem, e tentar encontrar soluções que simplifiquem as variáveis do problema.

Desta forma a geoestatística também tem outra limitação: **Os modelos geostatísticos requerem amostras realizadas em quantidade e qualidade adequada para gerar resultados satisfatórios.** Esta talvez seja uma das limitações mais difíceis de se conseguir abordar dentro da mineração. Para estimar de forma adequada precisa-se de amostras, e amostras são caras. Muitas empresas deixam de amostrar adequadamente seus depósitos minerais com finalidade de redução de custos, mas acabam por avaliar mal seus depósitos minerais, e consequentemente, obtém baixo lucro ou inviabilizam o uso sustentável dos recursos minerais. O termo qualidade, também é uma questão muito importante. Muitas vezes as amostragens realizadas possuem protocolos mal dimensionados. Alguns métodos de amostragem de jazidas também devem ser conduzidos de forma bem precisa para realizarem estimativas adequadas, mas apesar de ser uma das etapas mais importantes na mineração, as empresas muitas vezes colocam a tarefa nas mãos de profissionais pouco qualificados.

## 2.7 Questões éticas na avaliação de depósitos minerais

Avaliar depósitos minerais é uma atividade incerta, devido a natureza dos depósitos minerais, no entanto, não há justificativa para o mal uso das técnicas, nem ao mesmo para decisões arbitrárias que não envolvam decisões puramente lógicas ou racionais. Infelizmente o setor mineral acaba por ser alvo de pessoas com má conduta, por ser uma área de grandes riquezas. Esta não é, com certeza, a personalidade da grande maioria dos trabalhadores que se dedicam diariamente no setor mineral, mas pessoas acabam por utilizar a justificativa do "incerto" para vender depósitos minerais subvalorizados. Um avaliador de depósitos minerais deve realizar sua tarefa friamente, analisando a viabilidade do depósito independente se ele gerará riquezas ou não.

É importante também para os gestores e gerentes de minas entenderem a natureza do problema, e que as incertezas geológicas produzirão muitas vezes resultados diferentes dos pretendidos. A mineração trata do aproveitamento de recursos que são limitados pelo tempo geológico de sua criação. Enquanto o ser humano ainda não controlar o tempo, é indiscutível que temos de aproveitar os recursos minerais existentes da melhor forma que consigamos. A avaliação de depósitos minerais é o alicerce das decisões na mineração, por isso é impreterível que os processos sejam realizados de forma mais correta possível.

**Proposição 2.7.1** *Está nas mãos do avaliador de depósitos minerais a determinação das condições necessárias para a progressão da lavra. O desenvolvimento de seus projetos deve seguir sempre com conhecimento e idoneidade, pois é dele que deriva o trabalho de pessoas, o aproveitamento correto dos recursos minerais e da sociedade que aproveita estes recursos*

## 2.8 Alguns conceitos iniciais sobre jazidas minerais

Apresentamos nesta seção alguns dos principais conceitos de engenharia de minas, necessários para a realização de trabalhos de geoestatística e avaliação de depósitos no setor mineral. Apesar deste livro possuir foco na geoestatística, consideramos adequado entender conceitos gerais da mineração, que influenciam nas decisões tomadas pela avaliação dos depósitos.

### 2.8.1 Minério

A definição de minério talvez seja uma das mais importantes na produção mineral. A sua determinação permite o aproveitamento econômico dos recursos minerais, decidindo o que deve ou não ser lavrado e aproveitado. Segundo [Hustrulid et al. \[2006\]](#), a definição de minério pode ser considerada como:

**R** "Um agregado mineral com um ou mais sólidos minerais aos quais podem ser minerados, ou dos quais um ou mais produtos minerais podem ser extraídos com lucro". [Hustrulid et al. \[2006\]](#)

Isto significa que nem em todas as ocasiões um minério será extraído com a finalidade de se obter lucro pela venda. Em alguns casos, as questões econômicas da extração mineral podem ser contra intuitivas neste sentido, devido a políticas externas, estados de guerra, monopolização da produção, entre diversos outros fatores. Neste caso preferimos adotar o conceito de minério a partir do seu benefício, nem sempre ele sendo econômico. Definimos minério como:

**Definição 2.8.1 — Minério.** *Minério é todo agregado mineral ou fóssil cabível de aproveitamento técnico, que possibilita um benefício, seja ele econômico ou social, de forma a propiciar os interesses das diferentes componentes da sociedade, sejam elas a União, as forças sociais ou mineradores.*

Um exemplo bem característico de minérios explotados contra o senso econômico são os minerais radioativos, de monopólio da União. É de interesse estratégico de um país deter estes recursos capazes de produzir energia e armas, sendo muitas vezes gastos valores acima do valor do minério para sua extração.

### 2.8.2 Teor de corte e teor crítico

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, permite sua extração econômica. 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). O teor crítico, no entanto, representa o teor ao qual se delimita o limite entre prejuízo e lucro. [Rendu \[2014\]](#) define o teor de corte como:

**R** "O teor de corte geralmente é definido como a mínima quantidade de um produto de valor ou metal que em uma tonelada métrica deve conter para que este material seja enviado para a planta de beneficiamento" - [Rendu \[2014\]](#)

### 2.8.3 Continuidade

A continuidade é um termo derivado em toda a história da matemática e da ciência desde tempos remotos. Talvez uma das primeiras concepções da continuidade seja com o paradoxo de Zenão, que conta a história da corrida de Aquiles e a tartaruga. [Srivastava and Parker \[1989\]](#) demonstram o sentido da continuidade como:

**R** "Uma descrição da similaridade ou da dissimilaridade entre pares de valores com uma função de sua separação do vetor  $h$ " - [Srivastava and Parker \[1989\]](#)

Em outras palavras podemos dizer que a continuidade espacial é representada pela similaridade entre medidas que se localizam em regiões diferentes no espaço. Os fenômenos geológicos, neste caso, apresentam uma importante característica derivada de suas gêneses: Na maioria dos casos, medidas de propriedades realizadas mais próximas tendem a ser mais similares entre si do que medidas realizadas em grandes distâncias. Caracterizar a similaridade dos fenômenos geológicos é a chave para garantir que as estimativas e a caracterização da incerteza geológicas possam ser realizadas.

**Definição 2.8.2 — Continuidade espacial.** Definimos a continuidade espacial como a regularidade com que uma propriedade é medida em amostras aproximadas no espaço. Se as diferenças entre as amostras for pequena, dizemos que o material é contínuo ou similar. Quando o material é muito diferente de amostras pouco espaçadas dizemos que ele é discreto ou dissimilar. Na geoestatística definimos a continuidade a partir de uma direção do espaço, podendo ela se apresentar diferencialmente de acordo com a direção adotada.

### 2.8.4 Diluição

Segundo [Susaeta et al. \[2008\]](#) a diluição se refere ao estéril que não é separado do minério durante a operação da lavra. Este estéril é misturado com o minério e enviado para a usina de beneficiamento. Enquanto aumenta a quantidade de material enviado para a usina a diluição diminui o teor que deveria ser estimado e enviado para a usina corretamente. A estimativa de depósitos minerais é realizada desconsiderando os efeitos de produção e planejamento. Isto significa que os valores realmente lavrados não correspondem aos volumes planejados e induzem diferenças naturais da estimativa. O processo de comparação entre os valores reais obtidos na usina e os valores estimados pode ser definido como **aderência do planejamento de lavra**.

**Definição 2.8.3 — Aderência de lavra.** *Aderência do planejamento de lavra é todo o processo de comparação entre os valores estimados do depósito mineral e os obtidos durante a operação, seja durante a mineração, ou dos valores obtidos durante o beneficiamento mineral.*

As incertezas geológicas presentes no depósito mineral, ou as diferenças do planejamento da operação podem trazer discordâncias quanto os volumes e qualidade do material estimado e realmente lavrado, causando diluição do minério. Existem diferentes tipos de diluição durante a extração mineral. A **diluição interna** ocorre quando existem partes de estéril dentro do volume estimado do minério. Algumas vezes a amostragem pode não computar veios ou lentes de estéril dentro do bloco de decisão de lavra, dado que o volume das amostra é muito inferior ao volume das amostras. A **diluição externa** ocorre quando o planejamento mineral aborda parte do material não definido como minério, o que é comum nas regiões de contato do corpo geológico. Também há a chamada **diluição operacional**, que ocorre quando o desmonte de rochas realiza a fragmentação em regiões acima do planejado, chamado de *overbreak*, ou abaixo do planejado, chamado de *underbreak*.

### 2.8.5 Recursos e reservas minerais

A definição de recursos e reservas minerais são alternativas para publicidade de declarações públicas relativo às incertezas geológicas do depósito mineral. A CBRR (Comissão Brasileira de Recursos e Reservas) identifica a declaração pública como:

**Definição 2.8.4 — Declaração pública.** *Declarações públicas são preparadas para informar investidores ou potenciais investidores e seus conselheiros sobre os resultados da exploração, recursos minerais ou reservas minerais. Elas incluem, mas não se limitam, a relatórios anuais ou trimestrais das entidades, notas à im-*

*prensa, memorandos informativos, documentos técnicos, publicações em website e apresentações públicas.*

A partir de declarações públicas, as empresas podem indicar os volumes de metais e de massas estimados com base no conhecimento da incerteza geológica. Esta alternativa foi criada na década de 70, principalmente após o escândalo da empresa Bre-X, após constatado salgamento das minas de ouro em Busang na Indonésia. Definindo **Recursos** e **Reservas** minerais, o minerador classifica seus potenciais de produção segundo a incerteza geológica. A CBRR também define Recurso Mineral como:

**Definição 2.8.5 — Recurso Mineral.** *Um Recurso Mineral é uma concentração ou ocorrência de material sólido de interesse econômico dentro ou na superfície da crosta terrestre onde forma, teor ou qualidade e quantidade que apresentem perspectivas razoáveis de extração econômica.*

A definição de Recurso está ligada diretamente ao conhecimento da incerteza geológica. Os códigos de mineração não definem as técnicas necessárias para se definir os volumes de depósito de acordo com estas incertezas, apenas indicam que deve-se usar alguma técnica pertinente para isto. A responsabilidade desta definição cai diretamente à pessoa competente responsável pela auditoria. Estes Recursos minerais podem ser divididos em ordem crescente de confiabilidade geológica de acordo com as categorias de Inferido, Indicado e Medido. A CBRR também define Reserva Mineral como

**Definição 2.8.6 — Reserva Mineral.** *Uma Reserva Mineral é a parte economicamente lavrável de um Recurso Mineral Medido e/ou Indicado. Isso inclui diluição e perdas que podem ocorrer quando o material é lavrado ou extraído e é definido apropriadamente pelos estudos nos níveis de Pré-Viabilidade ou de Viabilidade que incluem aplicação de Fatores modificadores.*

Ou seja, para transformar um recurso em reserva mineral é necessário que se prove a viabilidade da extração do minério, seja ela econômica, social, ambiental ou política. Isto é realizado a partir dos fatores modificadores. A CBRR também define os fatores modificadores como:

**Definição 2.8.7 — Fatores Modificadores.** *Fatores Modificadores são considerações usadas para converter Recursos Minerais em Reservas Minerais. Esses incluem, mas não se limitam a considerações sobre: a lavra, o processamento, a metalurgia, a infraestrutura, a economicidade, o mercado, os aspectos legais, ambientais, sociais e governamentais*

As Reservas Minerais podem se dividir em provável, quando medida a partir de

um Recurso Indicado e, em algumas circunstâncias de um Recurso medido. A reserva provada é aquela que possui alta confiabilidade, representando recursos medidos. A figura 2.7 demonstra graficamente os resultados da exploração mineral em recursos e reservas minerais, também apresentando sua forma de conversão, de acordo com o conhecimento geológico e os fatores modificadores.

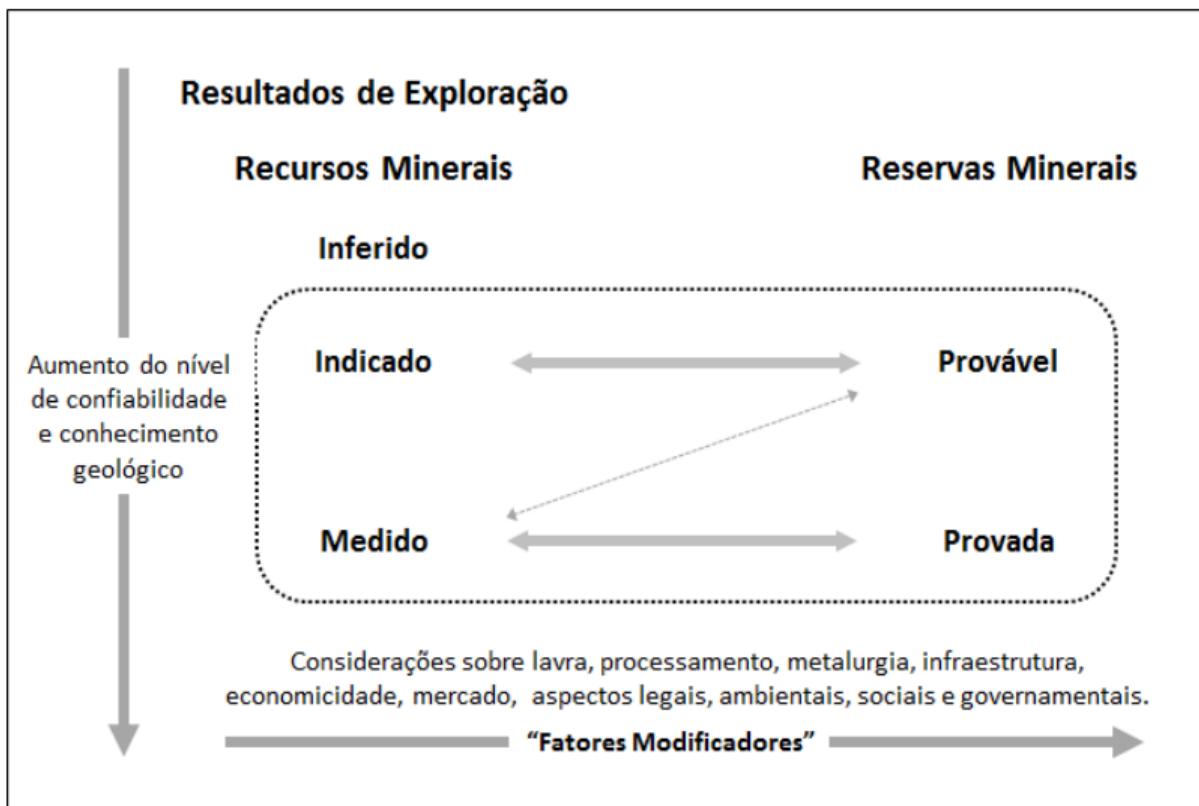


Figura 2.7: Figura demonstrando a classificação de jazidas em recursos e reservas. Linhas indicando a transição entre as classificações

### 2.8.6 Precisão e Exatidão

Uma das premissas utilizadas na geoestatística clássica é que os resultados das amostras obtidas é um valor fixo. Esta afirmação na maioria dos casos não é realista, pois as amostragens na mineração podem apresentar diferentes valores referentes aos erros de amostragem.

**Proposição 2.8.1** *Para os processos de geoestatística clássica, os valores das amostras georeferenciados são determinísticos, a medida que apresentam volume, posicionamento e propriedades constantes. Isto não se aplica a todos os métodos geoestatísticos como o KVME (Kriging with Measurement Error Variance) Delhomme [1978]*

Esta variação das amostras quanto ao valor esperado por elas pode ser definido por duas propriedades: **Exatidão** e **Precisão**. A Exatidão pode ser exemplificado como a proximidade de uma estimativa com a realidade, enquanto precisão é a medida da dispersão entorno de uma estimativa. Analogamente a precisão e a exatidão podem ser comparadas com um jogo de dardos como na figura (2.8), em que pretendemos atingir o centro do alvo. Quanto mais próximo forem os disparos do centro, melhor será a sua exatidão, e quanto mais próximos forem os disparos entre si, significa que são mais precisos. Disparos podem ser precisos, no entanto, não exatos. Disparos podem ser exatos por se localizarem em média próximos do centro, mas podem ser imprecisos se distanciarem entre si.

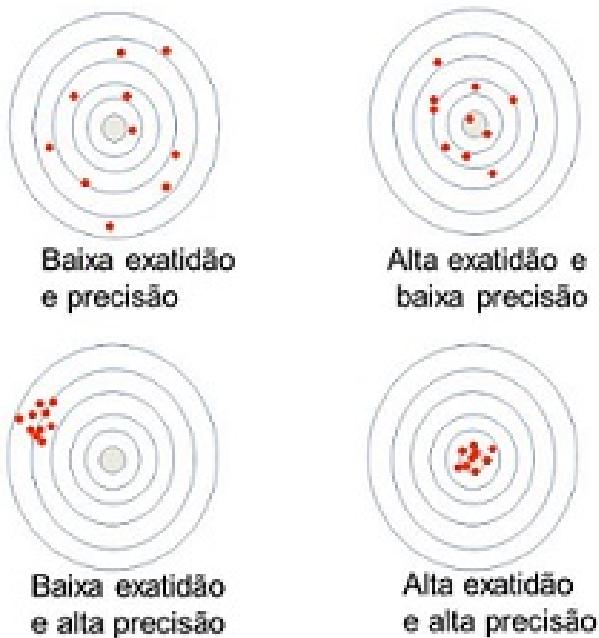


Figura 2.8: Figura demonstrando os conceitos de exatidão e precisão. O centro do alvo é o valor verdadeiro que pretende-se alcançar com os disparos. Disparos entorno do centro são considerados exatos. Disparos próximos aos outros são considerados precisos

A amostragem na mineração ainda sofre um outro problema, quanto a reproduzibilidade. Na verdade este é um problema para a maioria dos fenômenos espaciais, pois quando amostramos uma região não há como amostrar novamente, pois estas amostras geralmente são **destrutivas**. Além disso, ao amostrar em um local específico, uma amostra mesmo que próxima já se configura como uma amostra diferente. Os trabalhos do professor Gy [2012] invocam os principais conceitos e teorias a respeito da amostragem a granel, utilizada na mineração.

Eventualmente diversos fatores podem causar as variações e erros na amostra-

gem. 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.)

## 2.9 Conclusões

Neste capítulo inicial apresentamos os principais conceitos relacionados à geoestatística e ao planejamento de mina. Definimos o que é esta ciência que será abordada ao longo de todo o livro, o que ela pode realizar ou não, e sua importância dentro do contexto da mineração. Entendemos que a geoestatística é uma ferramenta para auxiliar na compreensão do desconhecido, e que é inerente ao empreendimento mineral, pois raras são as alternativas ao qual possuímos informação sistemática ao longo de todo o depósito mineral.

## 2.10 Exercícios

**Exercícios 2.1** Segundo a definição de Carrasco et al. [2008], sabemos que uma variável é aditiva se o seu valor médio é igual a média de seus valores. Discutimos ao longo do texto que as variáveis teor e conteúdo metálico são variáveis aditivas, capazes de serem utilizadas nos modelos clássicos que abordamos neste livro. Desta forma identifique variáveis na mineração que podem ser consideradas

aditivas ou não.



**Exercícios 2.2** 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?



**Exercícios 2.3** 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%.



**Exercícios 2.4** Os dados da tabela seguinte demonstram um conjunto de valores estimados e dados reais obtidos. Determine:

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







### 3. Variáveis aleatórias regionalizadas

*Todas as vezes que eu leio relatórios estatísticos, eu tento imaginar meu contemporâneo infeliz, a Pessoa Média, a quem, de acordo com estes relatórios, possui 0.66 filhos, 0.032 carros e 0.046 TVs.*

*Kato Lomb*

#### 3.1 Introdução ao capítulo

A geoestatística é uma ciência que se iniciou nos anos de 1950, com estudos de Krige [1960] na África do Sul a respeito de valores estimados em distribuições lognormais de ouro. Em 2012 o professor Daniel Krige recebeu a Ordem de Baobab, uma condecoração do presidente da África do Sul, pelas suas excepcionais contribuições para a economia, ciência, medicina, inovações tecnológicas e serviços comunitários. Durante seus 30 anos de idade, se tornou pioneiro no uso da estatística para avaliação de depósitos de ouro para um número limitado de furos de sondagem. As ideias do pesquisador foram fortemente abraçadas pela França após a tradução de seus artigos em língua nativa em 1995, o que gerou a fundação do centro de Geoestatística em Fontainebleau, corroborando para os estudos do professor George Matheron, e a criação da **teoria das variáveis regionalizadas**.

Este primeiro capítulo introduz a geoestatística a partir do seu objeto de estudo, as variáveis regionalizadas. Explicamos os principais conceitos abordados pela teoria clássica, e como eles se relacionam no entendimento dos fenômenos espacializados. Maiores informações podem ser encontradas nas obras de Matheron [1963] ou nos livros base de Isaaks and Srivastava [1989] e Goovaerts [1997]

## 3.2 Variáveis aleatórias

Alguns conceitos iniciais sobre estatística são necessários antes que possamos aprofundar os conceitos de geoestatística. Um dos principais conceitos utilizados para o entendimento de fenômenos aleatórios é o de **variável aleatória**.

**Definição 3.2.1 — Variável aleatória.** *Uma variável aleatória é uma função de um espaço amostral  $S$  nos números reais.* Casella and Berger [2010]

Imagine que tenhamos um saco com grandes quantidades de pedras coloridas vermelhas e azuis. Nosso espaço amostral seria portanto  $S = \{\text{pedras vermelhas, pedras azuis}\}$ . Se quisermos determinar uma variável aleatória que seja definida pela amostragem de duas pedras poderíamos ter o seguinte resultado  $Z = \{(\text{pedra vermelha, pedra azul}), (\text{pedra vermelha, pedra vermelha}), (\text{pedra azul, pedra azul})\}$ . Uma variável aleatória geralmente é definida a partir de uma letra maiúscula, enquanto uma realização, ou seja, um resultado desta variável aleatória é definido por uma letra minúscula. A figura 3.1 é um exemplo de uma variável aleatória, pois para cada valor possível dentro do espaço amostral de diferentes litologias é associado um valor inteiro.

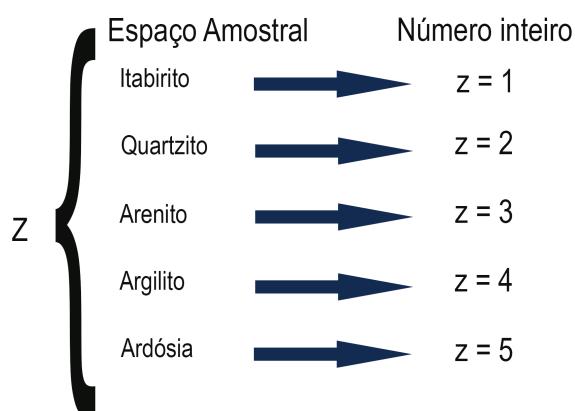


Figura 3.1: Exemplo de variável aleatória indicadora. Para cada possível valor de litologia do depósito é associado um valor inteiro.

Note, no entanto, que atribuir um valor para esta variável não significa dar uma maior importância ou uma menor importância para cada litotipo. Colocar um valor inteiro igual a 1 para o Itabirito não significa considerá-lo mais importante que as demais litologias. Neste caso dizemos que esta variável é **cardinal**, pois o valor associado de cada componente do espaço amostral a um valor inteiro não está diretamente ligado com sua importância, ao contrário de variáveis **ordinais** ao qual seu número associado é diretamente expresso pela sua importância.

As variáveis aleatórias são divididas geralmente em duas classes na geoestatística, considerando **variáveis aleatórias reais**, que podem apresentar valores dentro do conjunto de dados reais, ou **variáveis indicadoras**, quando consideramos que podem assumir valores inteiros. Exemplos de variáveis reais, por exemplo, são as de teores dos elementos metálicos, enquanto variáveis indicadoras são representadas pelas litologias presentes no depósito mineral.

Variáveis ditas **contínuas** são aquelas que possuem um espaço amostral infinito e **não contável**, geralmente representada por um conjunto de valores reais. Quando medimos teores, por exemplo, o resultado de uma amostra pode variar infinitamente dentro de um intervalo de 0% a 100%. Apesar desta limitação, o número de realizações que podem advir desta variação são infinitas, pois naturalmente o valor 5,6740 % é diferente do valor 5,6741 %, mesmo que muito próximos.

Em contrapartida, variáveis discretas são **contáveis**, mesmo que seu espaço amostral seja infinito. Se um subconjunto deste espaço amostral for considerado é possível conseguir definir para ele uma probabilidade. Variáveis discretas estão geralmente ligadas ao conjunto de números inteiros.

Para uma variável aleatória pode ser atribuído uma probabilidade ( $Pr$ ) de ocorrência para cada uma de suas realizações. A ideia de probabilidade mais básica está relacionada com a **frequência de ocorrência relativa de um evento**, ou também chamada de abordagem frequentista. Nossa variável aleatória demonstrada pela figura 3.1 pode ser associada a uma probabilidade de acordo com a figura 3.2, considerando a proporção de rochas de cada tipo dentro do domínio geológico estimado.

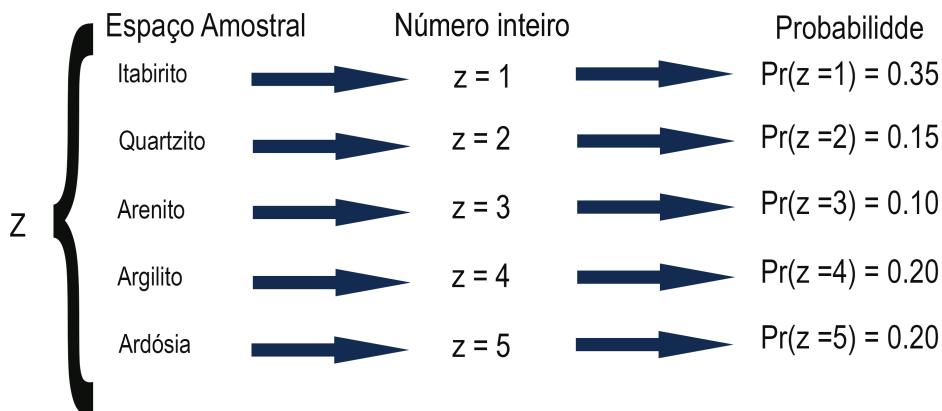


Figura 3.2: Exemplo de variável aleatória indicadora. Para cada possível valor de litologia do depósito é associado um valor inteiro. Uma probabilidade é atribuída para a frequência relativa de cada litologia.

Em outras palavras a probabilidade é semelhante a uma métrica de proporção das realizações de uma variável aleatória. Na verdade a probabilidade pode ser qualquer medida, desde que satisfaça os **Axiomas de Kolmogorov**.

1. A probabilidade de um evento é um número não negativo, dentro do intervalo  $[0,1]$ .
2. A probabilidade do espaço amostral é 1.
3. Se  $n$  eventos são mutuamente exclusivos, a probabilidade da união destes eventos é igual a soma das probabilidades individuais.

Os conceitos de probabilidade são estudados na matemática dentro da **teoria dos conjuntos** que é a base para a fundamentação da estatística. Para maiores informações da teoria base em probabilidade, axiomas de Kolmogorov e teoria dos conjuntos, aconselhamos ler as referências de [Alencar \[2014\]](#) e [FEITOSA et al. \[2011\]](#).

### 3.3 Função de distribuição acumulada - fda

Para cada elemento de uma variável indicadora podemos associar um valor de probabilidade, ou de frequência da apresentação deste elemento. Por exemplo se consideramos que um depósito mineral possui apenas dois tipos de rocha, podemos dizer que o tipo 1 representa 30% de frequência no depósito mineral, enquanto o tipo

2 apresenta 70% de frequência. Associar uma probabilidade para variáveis aleatórias indicadoras é intuitivo. No entanto, não conseguimos definir a probabilidade de um elemento para variáveis aleatórias reais contínuas, pois o espaço amostral é infinito. Não conseguimos associar, por exemplo, a probabilidade de um teor ser 5,67%. Neste caso utilizamos uma abordagem intervalar, associando a probabilidade a um intervalo de valores reais, logo é possível dizer que o depósito mineral possui probabilidade de 40% dos teores variarem de 5,67% a 9,32%. Uma função de distribuição acumulada é representada pela probabilidade de uma variável aleatória assumir um valor igual ou menor a um determinado limite. Definimos então a **Função de distribuição acumulada**  $F(z)$  tal como

$$F(z) = \Pr(Z \leq z) \quad (3.1)$$

A figura 3.3 indica a função de distribuição para variáveis contínuas e discretas. Em A) possuímos uma função discreta que pode assumir apenas valores inteiros de 1 a 8. Em B) possuímos uma função contínua de valores que se alteram no intervalo [1,8]

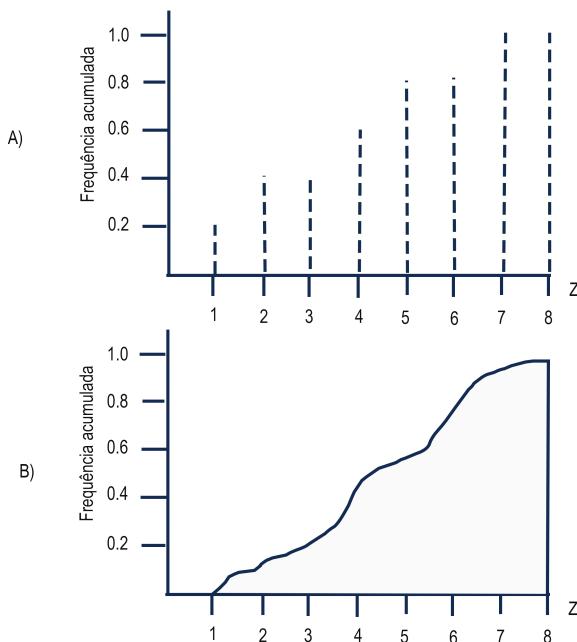


Figura 3.3: Função de distribuição acumulada - fda para variáveis discretas A) e contínuas B)

### 3.4 Função de densidade de probabilidade - fdp

No caso de distribuições contínuas, em que os valores podem ser determinados para qualquer valor dentro de um domínio real, é permitido utilizar princípios de cálculo para medir informações dessas distribuições. Como não podemos definir o valor da probabilidade em um ponto específico precisamos utilizar o conceito de probabilidade intervalar. A **Função de densidade de probabilidade-fdp** pode ser determinada como a probabilidade um valor assumir este valor dada uma variação infinitesimal.

$$f(z) = \lim_{\delta \rightarrow 0} [Pr(Z > z, Z < z + \delta)] \quad (3.2)$$

Logo a relação entre a função de distribuição acumulada e a função de densidade de probabilidades pode ser expressa por

$$F(z) = \int_{-\infty}^z f(z) dz \quad (3.3)$$

**Proposição 3.4.1** *Pode parecer que o valor da densidade de probabilidade seja equivalente ao valor da probabilidade assumindo uma realização  $z$  de variável aleatória  $Z$ , no entanto esta visão está errada! A ideia de probabilidade está diretamente ligada na ideia de frequência relativa de um evento. Valores de variáveis contínuas possuem espaço amostral incontável, o que significa que não conseguimos medir o seu tamanho, sendo ele infinito. Quantos seriam os possíveis resultados, por exemplo, de um teor de uma amostra apresentar? Se fosse possível associar uma probabilidade a um valor de uma variável aleatória real, todos estes valores seriam iguais a zero, pois sua frequência em nada representa na imensidão de valores possíveis. A função de densidade de probabilidade é na verdade uma medida de taxas de variação da probabilidade.*

### 3.5 Variáveis regionalizadas

Matheron [1963] , pai fundador da geoestatística, iniciou o conceito de **variável regionalizada**, para exemplificar os fenômenos espaciais. Quando um fenômeno exibe uma certa **estruturação espacial**, dizemos ele ser regionalizado. Os fenômenos geológicos, por exemplo, exibem estruturação característica na sua formação, o que significa que os corpos geológicos apresentam geometrias características de suas gêneses. A variável regionalizada  $z(x)$  denota um valor conhecido em um determinado ponto  $x$ , sendo apenas um resultado neutro puramente descritivo, sem

interpretação probabilística. Em outras palavras, a variável regionalizada é o valor real encontrado no depósito mineral para cada ponto  $x$  no espaço.

**Definição 3.5.1 — Variável Regionalizada.** *Uma variável regionalizada  $z(x)$  representa a medida de uma propriedade qualquer, seja ela o teor do elemento metálico, quantidade de metal ou acumulação, definida em um ponto  $x$  no espaço de coordenadas definidas*

É impossível para nós conhecer o valor real de  $z(x)$  para cada ponto no espaço, pois implicaria em muito mais que uma amostragem sistemática por todo o domínio do depósito mineral. Desta forma a variável regionalizada apresenta aspectos contraditórios, porém complementares para a definição do modelo geoestatístico:

- **Apresenta uma componente aleatória** onde não conseguimos amostrar ou definir a variável. Isto marca o aspecto irregular da variável. Seguindo a notação estatística, nos locais onde a variável regionalizada não é definida, denotamos  $Z(x)$  para informar que nestes locais ela assume aspecto de uma variável aleatória. Sendo  $\Omega$  o universo que pode ser composto a variável, definimos a variável aleatória em local desconhecido da variável regionalizada como  $Z(x) : \Omega \rightarrow \mathbb{R}$ .
- **Apresenta uma componente estruturada** nos locais onde é determinada, como por exemplo, pelos métodos de amostragem, representada pela própria forma  $z(x)$ , convencionalmente pela notação estatística como a realização da variável aleatória  $Z(x)$  no suporte  $x$ .

**Proposição 3.5.1** *Pode parecer um tanto estranho que algo possa assumir condições dicotômicas desta forma. Ao mesmo tempo que consideramos que algo existe e é determinístico, também consideramos que algo é aleatório e transitório. Na verdade as coisas são como sempre são, o que fazemos é assumir que em certos casos, não conseguimos definir algo, e em outro sabemos muito bem o que é. A aleatoriedade, na verdade, nunca existiu. Aleatoriedade é nosso princípio de humildade em não entendermos como os fenômenos ocorrem.*

A observação da variável regionalizada, não ocorre, no entanto em um ponto do espaço. Pontos são abstrações matemáticas de dimensão infinitesimal, uma condição geralmente para que possamos aplicar o princípio de continuidade dos modelos. As nossas observações são realizadas em amostras com volumes específicos e em grandes regiões que queremos estimar. Matheron [1963] apresenta os principais conceitos de domínio e suporte. Um domínio é uma região onde a variável regionalizada é diferente de zero. No nosso livro apresentamos o domínio das estimativas pela

notação  $D$ , enquanto os domínios de um bloco ou painel de lavra são apresentados por  $V$ .

**Definição 3.5.2 — Domínio.** *Domínio de uma variável regionalizada pode ser considerada qualquer região onde a variável apresenta valor diferente de zero. Por exemplo, a região da mina onde pretendemos estimar valores desconhecidos pode ser considerada como um domínio de estimativa  $D$ .*

Matheron [1963] apresenta também o conceito de suporte, sendo este relacionado com a capacidade de entendimento da variável regionalizada  $z(x)$ . De certa forma, é impossível conhecer o valor da variável regionalizada em um ponto  $x$ , pois o que detemos é o conhecimento da variável em um volume  $v$ , representando um testemunho de rocha, ou um fragmento de rocha.

**Definição 3.5.3 — Suporte.** *Suporte é o volume e forma  $v$  ao qual se detém o conhecimento da variável regionalizada  $z_v(X)$*

Em alguns casos, pela dimensão do domínio estimado em relação ao suporte, este é quase observado como um ponto. Imagine um fragmento de rocha de  $10\text{cm}^3$  e um painel a ser estimado de  $200\text{m}^3$ . A diferença de ordem de grandeza entre a amostra e o painel é gigantesca.

**Proposição 3.5.2** *Dizemos que do ponto de vista matemático é quase impossível definir a variável regionalizada  $z(x)$ , pois é quase impossível amostrar em um ponto. No entanto, esta é uma observação muito purista, que desconsidera os aspectos de engenharia. Em alguns casos uma amostra pode ser visualizada como uma realização da variável regionalizada  $z(x)$ , pois o volume da amostra é tão inferior ao domínio, que se torna praticamente uma dimensão pontual*

Uma das condições de aplicação da geoestatística clássica, que considera o uso de variáveis aditivas, é que o suporte das amostras utilizado nas estimativas deve ser o mesmo. Isto significa que o volume dos testemunhos utilizados para estimativa, amostras de canais, ou outros tipos de amostragens devem ter todos mesma forma, tamanho e volume. Esta é também outra questão impraticável, pois é impossível principalmente em rochas, obter regularidade nas amostras desta forma. Para contornar esta situação nos utilizamos os métodos chamados de **regularização**, que permitem criar amostras de mesmo tamanho.

Estas definições são as clássicas apresentadas pelo professor George Matheron em seus primeiros trabalhos sobre a teoria das variáveis aleatórias regionalizadas. Existe muita confusão entre diferentes autores para a representação destes conceitos de **suporte** e **domínio**, sendo muitas vezes o domínio do painel chamado de suporte do painel. De acordo com a definição de suporte, seria necessário conhecer o valor

real do painel, o que é impossível, sendo mais adequada a nomeclatura de domínio do painel. Estas divergências de conceituação não prejudicam o estudo da geoestatística como um todo, mas acabam por criar diferentes formas de notação e algumas vezes dificultam a leitura dos textos. O mais importante em se ter em mente é que este volume, seja do domínio ou do suporte, pode alterar os resultados das suas estimativas, na chamada **relação volume e variância**.

Esta ambiguidade da variável regionalizada permite tratamento de forma diferenciada segundo os objetivos de cada estudo. Podemos, ora tratar a variável regionalizada apenas como valores dispostos no espaço, ora dar um tratamento probabilístico para estes valores. Matheron [1963] aborda estes dois princípios como

- **Métodos transitivos** Considera a hipótese de estacionaridade, mas não implica em qualquer hipótese probabilística, sendo métodos apenas descritivos da variável regionalizada  $z(x)$ . Esta abordagem é utilizada principalmente na geoestatística clássica abordada neste livro. Faremos os cálculos geoestatísticos considerando apenas a descrição dos valores amostrados em uma determinada região, sem premissas sobre uma possível distribuição de probabilidades local.
- **Teoria intrínseca** Utiliza a interpretação probabilística da variável regionalizada  $Z(x)$ , também considerando hipóteses de estacionaridade. Esta metodologia é amplamente utilizada nos métodos considerados não-lineares e nas simulações geoestatísticas, em que se pretende determinar não apenas um valor esperado determinístico para um volume estimado, mas também uma distribuição de probabilidades.

## 3.6 Funções aleatórias

Como dissemos anteriormente a variável regionalizada possui uma componente tanto determinística, onde conhecemos os valores da variável, como uma componente aleatória, em locais onde se desconhece a propriedade de interesse. Este aspecto dicotômico é trocado por alguns autores ao usarem da **teoria intrínseca** e estabelecerem a variável aleatória sobre termos exclusivos de uma interpretação probabilística. Uma visão um pouco mais abstrata da variável aleatória é entender que sua componente determinística é apenas um resultado ou uma realização da variável aleatória naquele local, e que  $Z(x)$ , chamada em alguns casos de **função aleatória** é uma função que associa a qualquer ponto do espaço uma variável aleatória. A figura 3.4 demonstra o resultado de uma amostragem  $z(x = x_1)$  no ponto  $x_1$ .

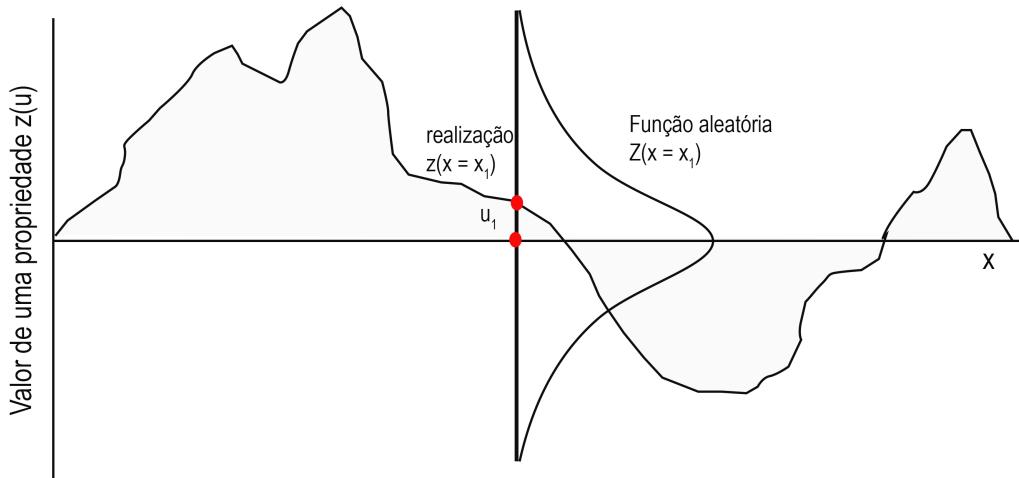


Figura 3.4: Demonstração do resultado amostrado  $z(x = x_1)$  como uma realização da função aleatória  $Z(x = x_1)$ . No ponto  $x_1$  o valor amostrado é apenas um resultado de uma função que desconhecemos, que associa uma distribuição de probabilidades naquele local.

Em muitos os casos não é possível conhecer esta função geradora do depósito mineral, apenas tomamos como hipótese que ela existe e é uma combinação de variáveis aleatórias em todo o espaço. Na geoestatística muitas vezes consideramos que esta função pode ser representada como uma combinação linear destas variáveis, chamada de **geoestatística linear**, ou **geoestatística clássica**. Ao tomarmos esta simplificação proposta pela teoria intrínseca, a demonstração das técnicas geoestatísticas se tornam bem mais fáceis, por isso, durante este texto, pretendemos utilizar o conceito da função aleatória em vez da forma tradicional da variável regionalizada proposta por Matheron.

**Definição 3.6.1 — Função aleatória.** *Uma função aleatória pode ser descrita como uma função que associa a cada ponto no espaço  $x$  uma variável aleatória  $Z(x = x_1)$ , sendo  $x_1$  o ponto de coordenadas especificado.*

Esta função aleatória é composta de uma amalgama de diversas variáveis aleatórias, cada uma em um ponto do espaço. A análise geoestatística destes valores permite decompormos esta função em duas componentes principais de acordo com os valores esperados de cada uma destas variáveis. O valor esperado tende a ser o de maior probabilidade de ocorrência em um determinado local. Desta forma podemos decompor a função aleatória em duas componentes principais, o **resíduo** e o valor de **tendência**. Por definição, a função aleatória pode ser expressa por  $Z(x) = R(x) + m(x)$ , sendo que os resíduos possuem média igual a zero.

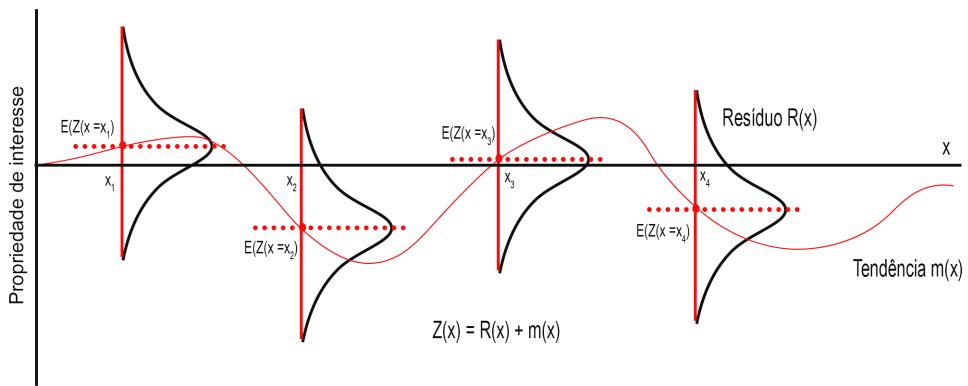


Figura 3.5: Decomposição da função aleatória a partir da determinação de sua tendência, indicada por  $m(x)$ , e o seu resíduo  $R(x)$ .

É comum na geoestatística assumirmos algumas hipóteses quanto a função aleatória. A **hipótese de estacionaridade de segunda ordem** afirma que o valor da tendência deve ser constante em todo o domínio considerado e o resíduo deve possuir variância constante para todo o domínio. Como desconhecemos a função aleatória, e nunca conseguimos determinar as variáveis aleatórias em cada ponto considerado, a hipótese de estacionaridade é sempre assumida, e nunca conseguimos comprová-la. Observe a série de números gerados na figura 3.6.

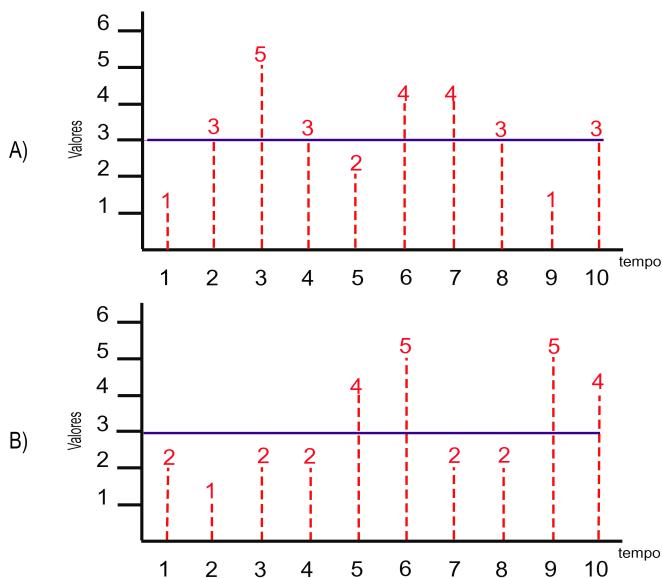


Figura 3.6: Série de números gerados em A e B. O valor médio destas séries é 2.9.

Ao observá-los, provavelmente você deve estar imaginando que foram feitos jogando-se dados na mesa. As séries A e B possuem média muito próxima do que seria de um dado de seis lados, e variam de 1 a 6. Na verdade, você está parcialmente certo,

eu gerei estes números a partir de dados. A diferença, no entanto, é que a série B foi gerada metade por um dado tetraédrico e metade por um dado cúbico, enquanto os dados da série A foram gerados apenas por um dado cúbico. Os valores médios reais que deveriam ser consideradas para este modelo são os representados na figura 3.7.

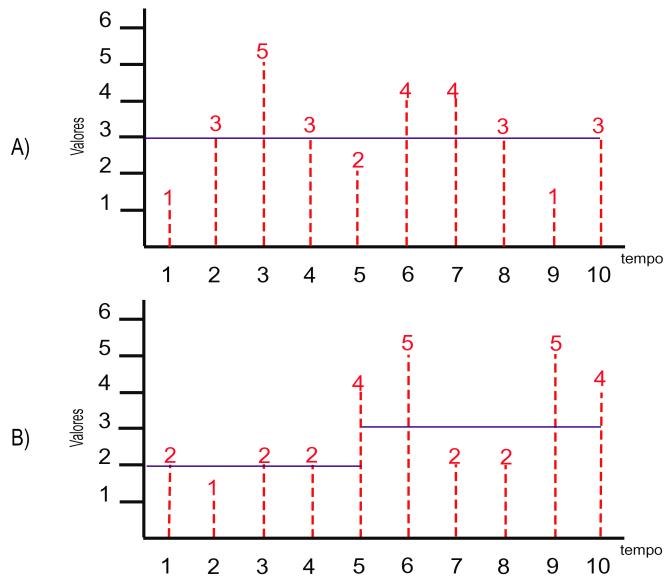


Figura 3.7: Série de números gerados. Em A) os dados foram gerados a partir de um dado tetraédrico, enquanto em B) foram gerados por um dado cúbico. Apesar dos valores médios globais serem idênticos, as médias locais são diferentes.

Apesar das médias globais serem exatamente as mesmas, as séries locais possuem distribuições distintas, com média e variância diferentes. Na verdade tanto a série B como série A poderiam ter sido geradas com o mesmo dado de seis lados. A diferença clara, quando consideramos cálculos **probabilísticos** com cálculos **estatísticos**, é que a probabilidade requer conhecimento sobre o fenômeno gerador, enquanto a estatística pretende inferir situações a partir das informações dos **dados**. Na verdade, a única informações que temos a todo momento no depósito mineral são amostras e informações indiretas como geofísica e geoquímica.



*A decisão de observar uma configuração particular dos dados como estacionário como o resultado de uma função aleatória estacionária está fortemente ligada com a decisão de que estas amostras podem ser unidas juntas. Nenhuma destas decisões pode ser checada quantitativamente, não são certas ou erradas e nenhuma prova das suas validades é possível. No entanto podem ser julgadas como apropriadas ou não. Isaaks and Srivastava [1989]*

Apesar de o fenômeno gerador ser completamente distinto para a metade dos

dados na série B, não é custoso unir estas diferentes distribuições sobre a mesma hipótese comum. Desta forma, assumir a estacionaridade neste caso é válido, dado que não conhecemos como estas informações foram construídas.

Em alguns casos, no entanto, não parece ser muito sábio adotar a hipótese de estacionaridade de segunda ordem. Observe a imagem da figura 3.8. A série é crescente com diferenças de valores iguais a 1 começando de um 1, 2, 3, 4, 5. Se você perguntasse para uma criança qual seria o próximo número na sequência ela diria 6. Se utilizássemos geoestatística para estimar o próximo número considerando a estacionaridade dos valores, o resultado seria 3.

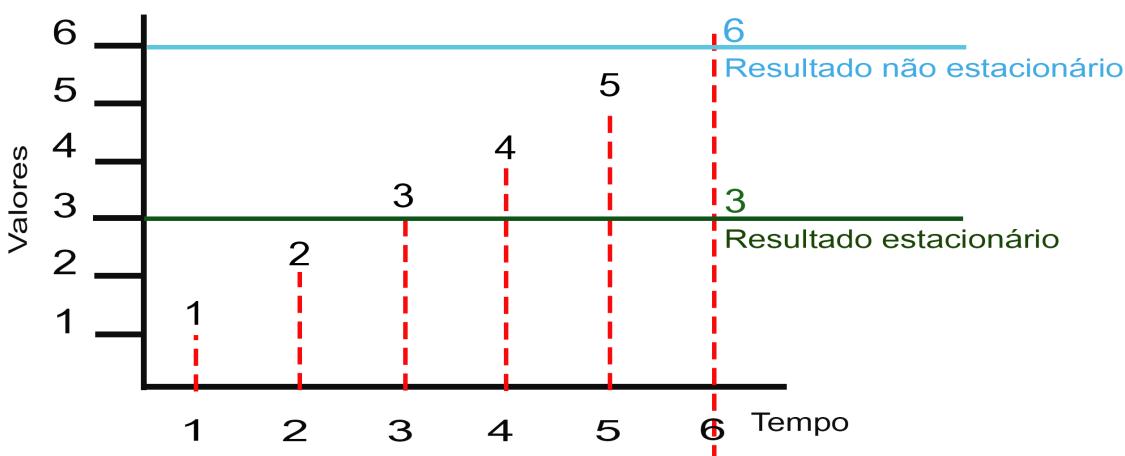


Figura 3.8: Série crescente de números. O próximo número da sequência a ser estimado considerando um modelo não estacionário seria 7, enquanto para o modelo estacionário seria de apenas 6

Não parece ser sábio adotar o número 3 neste caso. Se o fenômeno gerador desta série fosse realizado por um dado, seria tão equiprovável encontrarmos o número 3 ou o número 6 na próxima realização. No entanto a informação condicionada pela série parece nos instruir com clareza que existe este padrão a partir da observação indireta dos dados, nós desconhecemos como estes dados foram gerados. Utilizando a geoestatística nos reconhecemos que existe uma **estruturação** presente nesta sequência, que condicionalmente os dados gerados parecem seguir uma ordem, e que o próximo número gerado tente a apresentar uma variação talvez equivalente como ao dos anteriores, podendo ser 4 ou até mesmo 6.

**R** *Uma das ideias mais importantes na geoestatística é considerar que a função aleatória gera variáveis aleatórias condicionadas ao longo do espaço. A ideia de continuidade implica que qualquer informação próxima tende a ser mais parecida do que informações muito distantes. Isto é fisicamente plausível,*

*principalmente quando pensamos na geologia. As rochas que estão próximas tendem a possuir propriedades físico-químicas muito mais semelhantes do que quando consideramos uma distância muito grande. Por isso nos é intuitivo considerar o número estimado como 6 e não 3, pois ele representa um comportamento condicionado por medidas sucessivas, logo  $Pr\{Z(x_6) = 6|z(x_1) = 1, z(x_2) = 2, z(x_3) = 3, z(x_4) = 4, z(x_5) = 5\} > Pr\{Z(x_6) = 3|z(x_1) = 1, z(x_2) = 2, z(x_3) = 3, z(x_4) = 4, z(x_5) = 5\}$*

Estes casos também são chamados na geoestatística de **deriva**, ou seja, que existem mudanças graduais na tendência dos dados. Em alguns casos é bastante lógico na mineração considerar a deriva. A topografia, por exemplo, quando analisada em determinadas escalas e situações pode ser continuamente ascendente ou descendente. Neste caso descartar a hipótese de estacionaridade de segunda ordem é sábio.

**R** *O custo de aceitar o uso de um modelo inapropriado é que as propriedades estatísticas dos valores estimados divergirão de modelos homólogos Isaaks and Srivastava [1989]*

### 3.7 Hipótese de estacionaridade

Como visto anteriormente podemos realizar hipóteses a respeito da função aleatória, geradora dos fenômenos geoestatísticos. Estas hipóteses são decisões que não podem ser numericamente definidas, mas que em casos convém serem julgadas, para que as estimativas não retornem valores não condizentes com a realidade. A escolha de um tipo de estacionaridade significa que adotamos um critério que considere um conjunto de dados com um comportamento **homogêneo**. Uma das hipóteses utilizada pela geoestatística mais importantes, e que não constitui critério de escolha, é a chamada de **hipótese estrita**. Diferentemente da hipótese de estacionaridade de segunda ordem, esta é adotada automaticamente quando se opta por um método geoestatístico e não é passível de decisão. A principal ideia da estacionaridade estrita é que o fenômeno é homogêneo em uma mesma direção no espaço, sendo **invariante por translação**.

**R** *A ideia de areia em um jarro é uma boa imagem da estacionaridade de uma função aleatória em três dimensões, pelo menos enquanto a areia estiver bem ordenada (de outra forma se esta jarra vibrar, os grãos finos se depositarão na base, criando não estacionaridade vertical) Chiles and Delfiner [2009]*

Uma forma geométrica de pensarmos na hipótese estrita é pelo uso de fractais. Fractais são figuras geométricas autosimilares, em que cada um de seus componentes

carregam características da informação como um todo. A figura 3.9 representa um fractal. Estas formas autosimilares são muito comuns na natureza, seja no padrão desenhado por cristais de gelo, pela forma das plantas e principalmente nas rochas. A geologia em pequena escala muitas vezes é uma repetição que se traduz em grande escala.

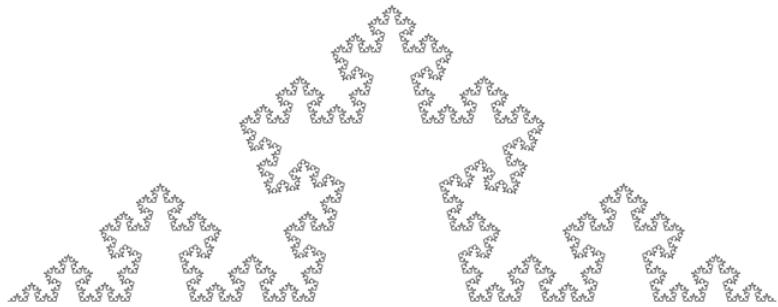


Figura 3.9: Fractal gerado a partir da repetição sistemática de estruturas cada vez menores.

**Proposição 3.7.1** *Todo o conhecimento humano somente advém do entendimento de padrões. As diferentes disciplinas, sejam elas humanas, biológicas ou exatas, apenas diferenciam quanto ao objeto de estudo. Não há diferença nenhuma entre um físico que entende padrões referentes ao movimento de planetas, um linguista que estuda o padrão de idiomas, um historiador que verifica padrões no tempo, ou um matemático que verifica o padrão das formas. A natureza também age desta forma, pois esperamos acordar no dia seguinte com o sol sobre as montanhas. Até mesmo dentro de fenômenos que parecem ser puramente aleatórios, podemos encontrar motivos pelos quais podemos entender padrões. Independente de fenômenos serem estacionários ou não estacionários, a geoestatística procura simplesmente estas formas no espaço, representações que apesar de não serem físicas, são mímicas da natureza da existência destes fenômenos*

Esta repetição de comportamentos em uma direção leva a seguinte definição matemática. Um fenômeno dito estacionário estrito significa que  $Pr\{Z(x_1) < z(x_1), Z(x_2) < z(x_2), \dots, Z(x_k) < z(x_k)\} = Pr\{Z(x_{1+h}) < z(x_{1+h}), Z(x_{2+h}) < z(x_{2+h}), \dots, Z(x_{k+h}) < z(x_{k+h})\}$ , sendo  $h$  um vetor de direção determinada. A hipótese de estacionariedade estrita significa que existe um grau de repetição no comportamento da variável ao longo de uma direção, no entanto, o fenômeno espacial pode apresentar deriva. Outra questão a ser abordada é o fato de que a adoção da estacionariedade é dependente da escala analisada. Um fenômeno considerado não estacionário pode assumir comportamento estacionário local. Observa a série de

dados representada pela figura 3.10. Quando analisado o comportamento global da função aleatória esta apresenta nitidamente uma tendência nos dados, no entanto, quando considerada uma escala menor do vetor  $h$ , este mesmo comportamento pode ser tomado como estacionário. Este fenômeno também é chamado de **quasi estacionário**.

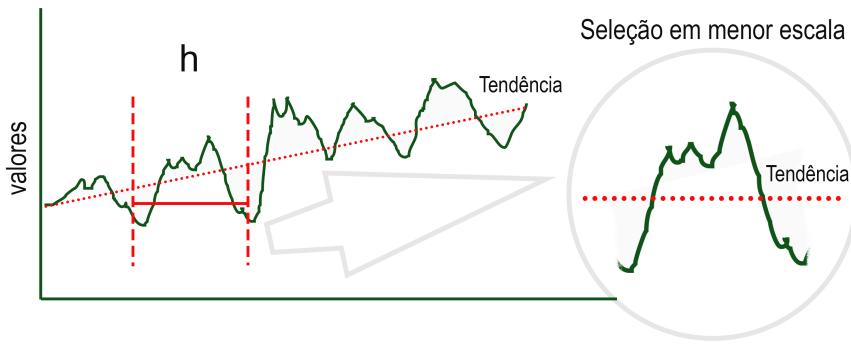


Figura 3.10: Comportamento analisado de uma série não estacionária quando analisada em um domínio global, e estacionária quando analisada em um domínio menor de comprimento  $h$ .

**Proposição 3.7.2** *Uma das maiores contribuições da geoestatística para a ciência talvez tenha sido a concepção de que os fenômenos podem ser dependentes da escala analisada. Dependendo da observação nossas hipóteses a respeito do fenômeno podem mudar. Isto é fisicamente compatível com a ideia da geologia. Analisar um depósito mineral em uma grande extensão de área, com toda a certeza é diferente quando observamos variações de tamanho centimétrico. A própria observação da Terra quando vista do espaço apresenta belos tons azuis e brancos, mas quando aproximamos a escala de uma região do tamanho de um país, notamos como nossa visão é diferente e muito mais variável.*

A hipótese de estacionaridade estrita é uma hipótese realizada sobre a característica do fenômeno, não dos resultados das amostras. A hipótese de estacionaridade de segunda ordem, no entanto, é uma hipótese relacionada com os **momentos estatísticos** do fenômeno. A principal ideia dos momentos estatísticos é que eles representam de alguma forma o resumo da distância entre os dados, desta forma quando pensamos na estacionaridade intrínseca, ou na estacionaridade de segunda ordem, pensamos na possível homogeneidade da distância entre os dados.

O conceito de **estacionaridade intrínseca**, desta forma, apresenta também outra forma de conceber esta homogeneidade, quando estabelecemos que uma variação  $Y_h(x) = Z(x + h) - Z(x)$  é estacionária de segunda ordem. Em outras palavras dizemos que existe homogeneidade quando consideramos a diferenças entre variáveis

aleatórias geradas pela função aleatória. Segundo [Chiles and Delfiner \[2009\]](#), se a hipótese de estacionaridade intrínseca pode ser considerada e não ocorre uma tendência, então o valor médio da função aleatória é constante, e o valor esperado de  $Y_h(x)$  é zero.

### 3.8 Momentos estatísticos

Como dito anteriormente, momentos estatísticos são representações da distância entre dados. As medidas de distância, podem ser por exemplo, medidas da tendência central dos dados ou medidas da dispersão destes dados. A figura 3.11 representa os conceitos de **tendência central** e de **dispersão**.

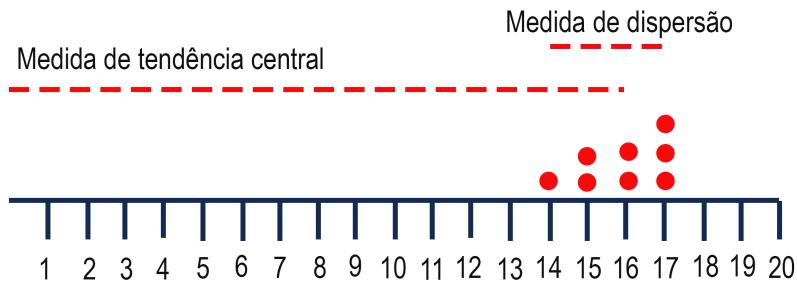


Figura 3.11: Exemplos de momentos estatísticos para um conjunto de dados. O valor médio representa o quanto distante está o centro dos dados, enquanto a dispersão apresenta quanto agregados estão estes dados.

A principal medida da distância do centro dos dados é chamada de **esperança matemática**, e pode ser representada para variáveis contínuas como:

$$E(Z) = \int_{z=-\infty}^{z=+\infty} z f(z) dz \quad (3.4)$$

No caso de variáveis discretas, podemos determinar a esperança matemática como

$$E(Z) = \sum_{i=-\infty}^{+\infty} Pr(z_i) z_i \quad (3.5)$$

Muitas vezes há confusões ao se dizer que a esperança matemática representa o valor mais provável que determinada variável pode possuir. No entanto, a esperança

matemática é simplesmente uma medida da distância do centro dos dados, sendo que este centro pode ser pouco provável ou nem mesmo existir. Observe a figura 3.12. A esperança matemática neste caso representa o centro de uma distribuição com probabilidade muito baixa de ocorrência.

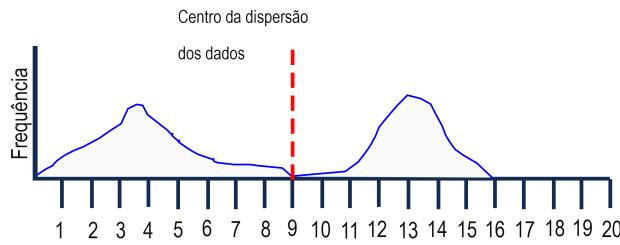


Figura 3.12: Exemplo da esperança matemática de uma distribuição multimodal. O valor da probabilidade para este centro de dispersão é praticamente nulo.

O que de fato ocorre é que os centros de dispersão dos dados da maioria dos problemas de engenharia não são multimodais, ou seja, apresentam vários picos nas distribuições de densidade de probabilidade com na figura. Neste caso a esperança matemática pode representar os valores mais prováveis de ocorrência da dispersão dos dados.

Pela definição da esperança matemática, algumas propriedades podem ser diretamente derivadas. A multiplicação da variável aleatória  $Z$  por um valor constante  $c$ , implica na seguinte condição.

$$E(cZ) = cE(Z) \quad (3.6)$$

O valor esperado de uma variável aleatória constante pode ser relacionada pela seguinte propriedade

$$E(c) = c \quad (3.7)$$

A demonstração, na verdade é muito simples, já que advém da própria definição de esperança matemática

*Demonstração.* Valor esperado de uma constante é igual a ela mesma

$$\begin{aligned} E(c) &= \sum_{i=-\infty}^{+\infty} Pr(c)c \\ E(c) &= c \sum_{i=-\infty}^{+\infty} Pr(c) \\ \text{Como: } \sum_{i=-\infty}^{+\infty} Pr(c) &= 1 \\ E(c) &= c \end{aligned}$$

■

A partir da definição da esperança matemática como uma medida de centralidade da distribuição dos dados, podemos derivar outros momentos representando diferentes distâncias desta distribuição. Os diferentes momentos matemáticos podem caracterizar diferentes distâncias relativas à **centralidade**, **dispersão**, **assimetria**, **forma**. A geoestatística, na verdade, foca sua análise principalmente nos momentos de primeira e segunda ordem.

**R** *Em aplicações da mineração, a lei de probabilidades espacial nunca é requerida, principalmente porque os dois primeiros momentos da função são suficientes para providenciar uma solução aproximada para muitos problemas encontrados Journel and Huijbregts [1978]*

Outro momento estatístico importante é a variância, definida como o momento de segunda ordem centrado. A variância pode ser considerada como uma medida de dispersão, demonstrada por

$$Var(Z) = E(Z - E(Z))^2 \quad (3.8)$$

Esta forma tradicional da variância pode ser substituída por outra representação a partir de

$$Var(Z) = E(Z^2) - E(Z)^2 \quad (3.9)$$

A prova desta relação também é facilmente demonstrada a partir das propriedades da esperança matemática e pela definição da variância.

*Demonstração.* Relação entre as equações 3.8 e 3.9

$$\begin{aligned}Var(Z) &= E(Z - E(Z))^2 \\Var(Z) &= E(Z^2 - 2ZE(Z) + E(Z)^2) \\Var(Z) &= E(Z^2) - E(2ZE(Z)) + E(Z)^2 \\Var(Z) &= E(Z^2) - 2E(Z)E(Z) + E(Z)^2 \\Var(Z) &= E(Z^2) - 2E(Z)^2 + E(Z)^2 \\Var(Z) &= E(Z^2) - E(Z)^2\end{aligned}$$

■

Os momentos estatísticos de primeira e segunda ordem, representados pela esperança matemática e pela variância representam medidas tomadas de uma única variável aleatória. Para relacionar diferentes variáveis aleatórias utilizamos comumente a **covariância**, esta representada pela similaridade entre duas variáveis aleatórias. Considere as variáveis aleatórias  $Z$  e  $Y$ . Podemos representar a covariância pela seguinte relação

$$Cov(Z, Y) = E((Z - E(Z))(Y - E(Y))) \quad (3.10)$$

Se as variáveis  $Z$  e  $Y$  apresentam médias idênticas iguais a  $m$ , então a covariância pode ser representada por

$$Cov(Z, Y) = E(ZY) - m^2 \quad (3.11)$$

A prova desta relação pode ser facilmente obtida

*Demonstração.* Relação da Covariância considerando médias idênticas iguais a  $m$

$$\begin{aligned}Como: E(Z) &= E(Y) = m \\Cov(Z, Y) &= E((Z - m)(Y - m)) \\Cov(Z, Y) &= E(ZY - Zm - Ym + m^2) \\Cov(Z, Y) &= E(ZY) - E(Zm) - E(Ym) + E(m^2) \\Cov(Z, Y) &= E(ZY) - mE(Z) - mE(Y) + E(m^2) \\Cov(Z, Y) &= E(ZY) - m^2 - m^2 + m^2 \\Cov(Z, Y) &= E(ZY) - m^2\end{aligned}$$

Se as variáveis Y e Z forem idênticas, a covariância entre as duas variáveis aleatórias é equivalente a variância. A prova pode ser demonstrada por

*Demonstração.* Prova de que a covariância é idêntica a variância para  $Z=Y$

$$\text{Como : } Y = Z \rightarrow Cov(Z, Y) = Cov(Z, Z)$$

$$C(Z, Z) = E((Z - E(Z))(Z - E(Z)))$$

$$C(Z, Z) = E(Z - E(Z))^2$$

$$C(Z, Z) = Var(Z) \vee Var(Y)$$

## 3.9 Ergocidade

A ergocidade é uma das propriedades mais importantes da função aleatória. A ideia é de que cada vez ao qual analisamos um volume maior no espaço, a tendência é que o valor médio deste volume se aproxime cada vez mais do valor médio do fenômeno. Matematicamente podemos definir a propriedade da ergocidade como

$$\lim_{V \rightarrow \infty} \frac{1}{|V|} \int_{x \in V} Z(x) dx = m \quad (3.12)$$

Em que  $|V|$  é o volume considerado e  $m$  o valor da média do fenômeno.

**Definição 3.9.1 — Ergocidade.** A Ergocidade pode ser caracterizada como a propriedade da função aleatória de convergência dos valores médios se aproxime a um valor constante  $m$ , de acordo com um domínio  $V$  considerado.

Alguns fenômenos tendem a apresentar dispersões infinitas, crescentes de acordo com o desenvolvimento da função aleatória. Estes fenômenos podem apresentar dispersão infinita, tal como o fenômeno de movimento browniano.

## 3.10 Homocedasticidade e heterocedasticidade

Além do comportamento da estacionariedade dos valores médios da função aleatória, também é importante qualificar os fenômenos geoestatísticos a partir do comportamento da variância. A figura 3.13, por exemplo, demonstra um comportamento crescente da variância de acordo com o desenvolvimento da série.

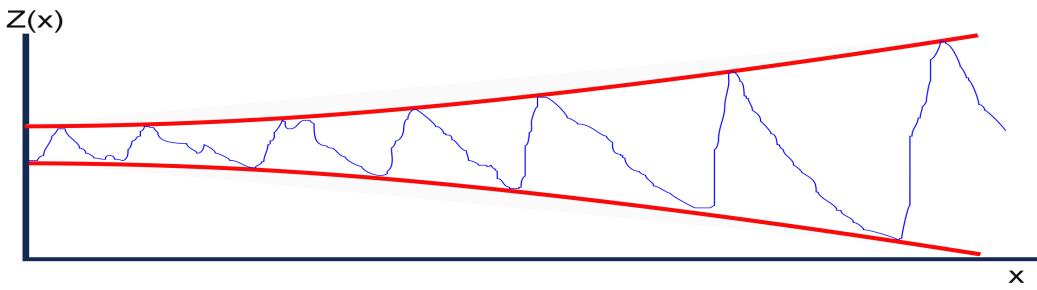


Figura 3.13: Fenômeno de heterocedasticidade representando variabilidade infinita da função aleatória

Estes fenômenos são chamados de **heterocedásticos**, e aumentam a variabilidade de acordo com o incremento da direção. Fenômenos constantes quanto a dispersão local são chamados de **homocedásticos**

**Definição 3.10.1 — homocedasticidade.** A hipótese de homocedasticidade, igual (*homo*) dispersão (*scedasticidade*), implica que a variância da função aleatória é constante para todo e qualquer ponto  $x$  representado no domínio  $D$ . A heterocedasticidade, no entanto, implica que a variância aumenta ao longo da função aleatória.

### 3.11 Relação Volume Variância

Na geoestatística, a variável aleatória se manifesta em todos os pontos no espaço. No entanto, nem sempre é possível reconhecer a variável em um suporte pontual, e para fins de engenharia precisamos entender a variável aleatória dentro de domínios específicos, sejam eles os domínios da amostra, na unidade seletiva de lavra, ou dentro de um domínio de estimativa. Na geoestatística clássica, apenas se determina os **valores esperados** destas variáveis dentro de um domínio, principalmente os de primeira e segunda ordem, não se importando com o reconhecimento das distribuições locais.

O processo de se determinar volumes esperados dentro de um domínio é chamado na geoestatística de **regularização**.

**R** *Muito raramente, em prática, o valor dos dados pontuais  $z(x)$  estão disponíveis. Mais comumente o valor dos dados  $z_v(x)$  em um certo suporte  $v(x)$  estão disponíveis, como por exemplo uma amostra de testemunho, ou mais genericamente o volume de uma amostra. O valor médio  $z_v(x)$  é chamado de regularização das variáveis pontuais  $z(y)$  dentro do domínio  $v(x)$  Journel and Huijbregts [1978]*

A regularização permite com que medidas realizadas dentro de um domínio estipulado possuam mesmo volume, permitindo com que suas propriedades sejam compatíveis para fins de estimativa na geoestatística.

Assumindo que a função aletória é contínua, e que uma combinação linear de variáveis aleatórias pode ser expressa, podemos definir um valor regularizado em um espaço amostral, definido pelo seu suporte. Considere  $v$  como o suporte amostral, logo o seu valor regularizado pode ser descrito como

$$Z_v = \frac{1}{|v|} \int_{x \in v} Z(x) dx \quad (3.13)$$

Da mesma forma o valor regularizado dentro da unidade seletiva de lavra pode ser definido por

$$Z_V = \frac{1}{|V|} \int_{x \in V} Z(x) dx \quad (3.14)$$

Em último caso podemos definir o valor médio dentro do domínio de estimativa

$$Z_D = \frac{1}{|D|} \int_{x \in D} Z(x) dx = m \quad (3.15)$$

Em que  $m$  é o valor esperado do fenômeno considerado, e constante, se considerada a propriedade da ergocidade. Considerando os diferentes domínios  $v$ ,  $V$ , e  $D$ , podemos determinar três diferentes relações  $(v|V)$ ,  $(v|D)$  ou  $(V|D)$ , representadas pelas relações entre amostras e unidade seletiva de lavra, amostras e domínio de estimativa e unidades seletivas de lavra e domínio de estimativa.

Para indicarmos a variabilidade ao qual os valores amostras em um domínio estão dispersos quanto um valor de referência estimado, utilizamos uma estatística chamada de **variância de dispersão**, denotada pela letra  $D^2$ . A ideia da variância está diretamente associada ao conceito de **entropia**, ou grau de desorganização.

**Proposição 3.11.1** *A variância de dispersão é uma das medidas mais importantes na geoestatística e está associada ao conceito de entropia, ou de desorganização dos dados. Quando você considera, por exemplo, a variabilidade de um pixel de uma foto em relação ao seu valor médio, com toda a certeza este será mais disperso que valores médios de partes do corpo na foto, como rostos e mãos, em relação a este valor central. A ideia de que nosso conhecimento sobre um fenômeno pode ser afetado pela dispersão da informação é essencial, principalmente nas técnicas de*

*mudança de suporte que serão vistas futuramente.*

A **variância de dispersão** é portanto uma medida da variabilidade entre estes domínios, considerando os valores regularizados. A variância de dispersão amostra e domínio de estimativa pode ser definida por

$$D^2(v|V) = \frac{1}{N} \sum_{i \in V} [Z_{v_i} - Z_V]^2 \quad (3.16)$$

Sendo  $N$  o número de pontos amostrais regularizados de suporte  $v$  dentro do domínio estimado  $V$ . Se considerarmos o suporte  $(.)$  como o suporte pontual, podemos definir a variância de dispersão ponto amostra por

$$D^2(.|v) = \frac{1}{|v|} \sum_{x \in v} [Z(x) - Z_v]^2 \quad (3.17)$$

Em que  $|v|$  é o volume constituído pelo suporte amostral  $v$  e todos os seus pontos internos. E a variância de dispersão ponto e domínio estimado por

$$D^2(.|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x) - Z_V]^2 \quad (3.18)$$

Em que  $|V|$  é o volume constituído pelo suporte amostral  $V$  e todos os seus pontos internos. Uma das relações importantes da variância de dispersão pode ser determinada pela diferença entre variâncias de dois suportes, tal como

*Demonstração.* Prova da relação da variância de dispersão entre ponto e bloco estimado como a diferença entre a variância do fenômeno e da variância do bloco

estimado.

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x) - Z_V]^2$$

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x)^2 - 2Z(x)Z_V + Z_V^2]$$

como  $\sum_{x \in V} Z(x)Z_V = Z_V^2$ , tal que  $Z_V = constante$

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} [Z(x)^2 - Z_V^2]$$

$$D^2(\cdot|V) = \frac{1}{|V|} \sum_{x \in V} ([Z(x)^2 - m^2] - [Z_V^2 - m^2])$$

Pela hipótese de estacionaridade de segunda ordem:

$$\frac{1}{|V|} \sum_{x \in V} [Z(x)^2 - m^2] = Var(Z(x)) = s^2(\cdot|.) , \text{ e}$$

$$[Z_V^2 - m^2] = Var(Z(V)) = s^2(V|V) , \text{ logo}$$

$$D^2(\cdot|V) = D^2(\cdot|.) - D^2(V|V)$$

■

Analogamente as relações  $s^2(\cdot|v) = s^2(\cdot|.) - s^2(v|v)$  e  $s^2(v|V) = s^2(V|V) - s^2(v|v)$  podem ser derivadas. Podemos encontrar então a seguinte identidade

$$s^2(\cdot|V) = s^2(\cdot|v) + s^2(v|V) \tag{3.19}$$

Esta também é chamada de **relação de krige** ou relação da aditividade de variâncias de krige. Quando consideramos a dispersão de valores de uma variável em domínios maiores como  $V$ , esta tende a ser maior que consideramos no suporte amostral  $v$ . Este princípio também é chamado de **volume e variância**, ou seja, quanto maior for a diferença entre os suportes amostrais e o domínio de estimativa, menor será nossa acurácia nestas previsões. A ideia da variabilidade de acordo com a mudança do volume estimado ou do suporte amostral está diretamente associada à definição de uma imagem, no conceito da geoestatística. Observe a figura 3.14. Em A) possuímos os valores exatos do fenômeno estudado. Podemos notar que os resultados são uma representação fiel de uma representação física, podem ser realmente consideradas um "mapa" dos valores distribuídos no espaço. Em B) verificamos os valores estimados, que se apresentam de forma pixelada e não apresentam

uma definição adequada do problema. No entanto, cada bloco estimado no mapa B) guarda uma correlação alta com os valores médios tomados da região no mapa A). Dizemos que as estimativas geoestatísticas não são uma ferramenta boa para produzir "mapas", já que estes são reproduções fidedignas dos fenômenos espaciais, mas o valor esperado dentro de um bloco em B) tende a ser cada vez mais próximo do valor médio real na região quanto menor for a definição e maior o tamanho do bloco.

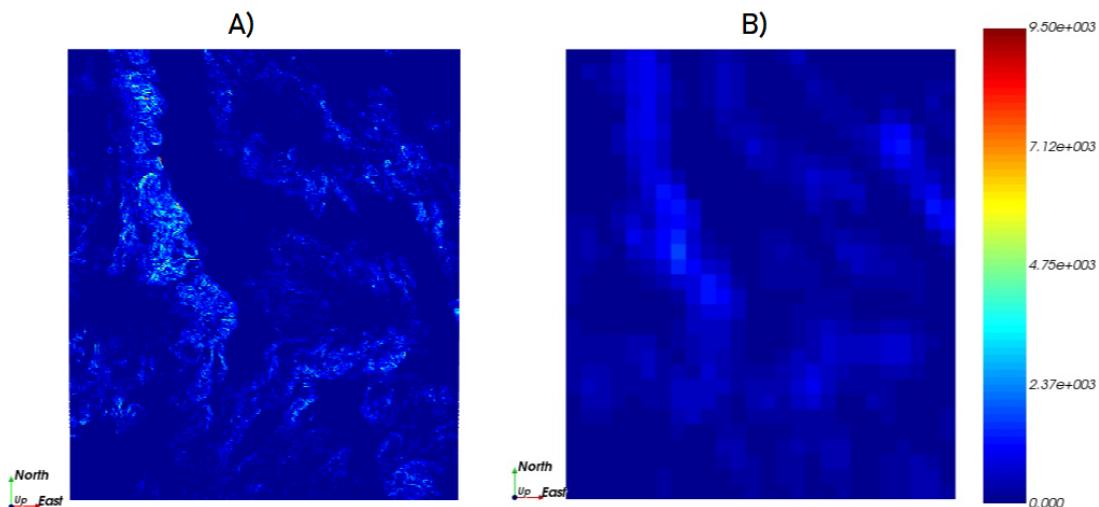


Figura 3.14: Relação do conceito de volume e variância apresentado em imagens. Em A) possuímos o valor exaustivo de um banco de dados, enquanto em B) apresentamos os valores krigados. É possível notar que as estimativas não reproduzem as feições naturais do fenômeno, no entanto, cada bloco estimado é

## 3.12 Conclusões

Neste capítulo aprendemos um pouco sobre a teoria das variáveis regionalizadas, um conceito determinado na década de 70 pelo professor George Matheron, e que evoluiu ao longo do tempo, facilitando o estudo de variáveis georeferenciadas. Estes conceitos iniciais são abstratos, porém poderosos, pois permitem constituir as bases de hipóteses utilizadas nos modelos geoestatísticos.

## 3.13 Exercícios

**Exercícios 3.1** 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. ■

**Exercícios 3.2** 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 ■

**Exercícios 3.3** 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% ■

**Exercícios 3.4** 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. ■

**Exercícios 3.5** Prove que a covariância de duas variáveis aleatórias independentes seja igual a zero. Dica.: Tome o valor de  $E(XY) = E(X)E(Y)$  ■





## 4. Estatística univariada

*Estatística: a ciência que diz que se eu comi um frango e tu não comeste nenhum, teremos comido, em média, meio frango cada um.*

*Pitigrilli*

### 4.1 Introdução

As avaliações geoestatísticas geralmente se iniciam com uma avaliação global das amostras. Nesta primeira etapa, o objetivo principal é **descrever** e **inferir** informações sobre o comportamento geral das amostras. A chamada **estatística descritiva** representa o conjunto de técnicas necessárias para resumir informações da realidade observada das amostras, usando formas numéricas ou gráficas para caracterizá-las. Já a chamada estatística inferencial ocupa em tomar inferências da população de dados a partir de informações das amostras. O estudo sistemático das variáveis em termos globais não representam o fenômeno estudado, mas partem do ponto de vista necessário para o início da pesquisa, podendo avaliar inconsistências nos dados e possíveis comportamentos que possam indicar situações favoráveis ou desfavoráveis na análise espacial.

**Proposição 4.1.1** *Usualmente, os sistemas aos quais estudamos não podem ser isolados em variáveis discretas e independentes. Estes fatores influenciam os primeiros*

*passos da pesquisa, em como e onde coletar espécies ou observações - Borradaile [2013]*

A palavra estatística, non entanto, apresenta duplo sentido. Pode representar a **teoria estatística** ou as medidas realizadas pelos dados. Alguns conceitos iniciais são de extrema importância quando consideramos o uso da estatística clássica univariada

**Definição 4.1.1 — População.** *conjunto de elementos que tem pelo menos uma característica em comum. No caso da geoestatística a população pode ser considerada analogamente ao conjunto possível de todas as realizações em um domínio geológico considerado*

**Definição 4.1.2 — Amostra.** *Amostra pode ser considerada como um subconjunto de elementos de uma população. Existem diferentes tipos de amostragens na mineração, como sondagens diamantadas, amostragens de canal, medições de nível freático, etc.*

Em muitos casos é comum representar este conjunto de dados por tabelas. Sumários que caracterizam as informações de cada subconjunto de amostras no espaço. Muitos softwares de geoestatística e planejamento mineral caracterizam os furos de sondagem a partir de dois ou três arquivos. Geralmente o primeiro arquivo consta uma tabela sobre o posicionamento da boca dos furos na superfície, caracterizando seu posicionamento espacial em um plano cartesiano  $\langle x,y,z \rangle$ . O segundo arquivo geralmente representa a direção dos furos e o comprimento realizado em cada manobra. E um terceiro arquivo geralmente apresenta as propriedades medidas em cada manobra realizada do testemunho.

Uma questão importante a ser considerada nas estatísticas descritivas é sua capacidade de resumo da informação. Estatísticas numéricas são uma alternativa importante para formar concepções que auxiliam na tomada de decisão, mas ao mesmo tempo reduzem a sensibilidade sobre outras questões dos dados. Imagine a figura 4.1. Temos duas fontes de temperatura equidistantes, uma com  $270^\circ$  e outra a  $-270^\circ$ . Apesar da diferença abrupta de temperaturas, a temperatura média da parede entre elas é apenas  $0^\circ$ . Um ser humano conseguiria sobreviver facilmente se ocupasse apenas o espaço entre estas duas fontes de temperatura, mas morreria se afastasse delas.

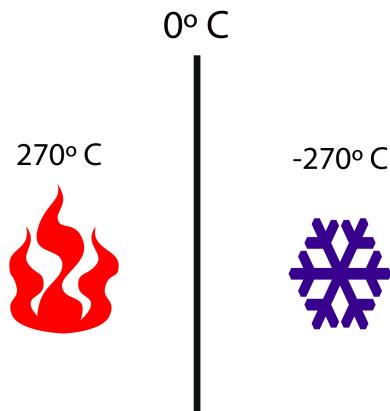


Figura 4.1: Duas fontes de temperatura equidistantes, uma quente e outra fria. A temperatura média da parede que separa estas fontes é igual a média das temperaturas, o que não representa toda a complexidade do fenômeno.

Apesar da média ser uma medida muito útil para ser utilizada na descrição dos dados, utilizada sozinha pode gerar interpretações erradas sobre o problema. Convenciona-se que o uso de estatísticas descritivas deve ser múltiplo, optando por utilizar não apenas uma, mas diferentes técnicas de avaliação.

O uso de estatísticas descritivas permite em muitos casos

1. Avaliar se as proporções globais possam estar acima do cut-off esperado
2. Identificar a facilidade da aplicação dos métodos clássicos de acordo com as distribuições de frequência
3. Auxiliar no dimensionamento de malhas de sondagem principalmente nas etapas iniciais (*greenfield*) na mineração
4. Identificar a possibilidade da divisão de domínios se apresentadas frequências multimodais

**Proposição 4.1.2** *Estatísticas univariadas são a primeira alternativa para analisar dados. Quando as amostras ainda são escassas, principalmente nas fases iniciais da pesquisa mineral, estas ferramentas são extremamente úteis para avaliarem de forma genérica os resultados das campanhas. Se bons resultados podem ser gerados a partir de estatísticas univariadas, a confiança no projeto aumenta suas perspectivas, no entanto, se os dados demonstrarem condições pobre das estatísticas, ainda podemos apostar em uma melhor avaliação do depósito*

É importante salientar que informações a partir de estimativa e interpolação não podem gerar dados além dos limites estipulados pela estatística descritiva univariada. 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.

## 4.2 Estatísticas pontuais

Como dito anteriormente, o conceito estatística pode ser dúvida, ao mesmo tempo que enfoca na 'teoria estatística' ou em **funções aplicadas em dados**. Quando estas funções forem aplicadas em todos os dados de um universo são chamadas de **parâmetros**. Qualquer função realizada a partir de dados pode ser considerada uma estatística, ou um **estimador**, no entanto, algumas delas são mais usuais, por conseguirem a partir de dados aproximar as estatísticas de seus respectivos parâmetros.

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.

**Definição 4.2.1 — Estatísticas pontuais.** *Estatísticas pontuais são funções realizadas a partir dos dados para calcular valores que representam propriedades do conjunto. Dentre as categorias mais conhecidas possuímos medidas de centralidade, dispersão, assimetria, achatamento*

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. Uma das ferramentas utilizadas para entender estas estatísticas visualmente também é conhecida como faces de Chernoff, como demonstrado na figura 4.2

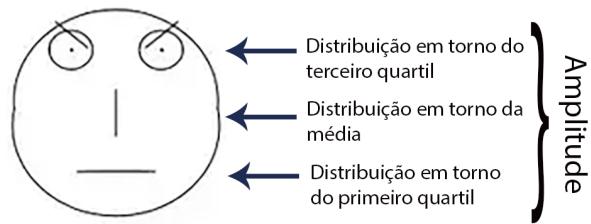


Figura 4.2: Exemplo das faces de Chernoff e as características das estatísticas com estruturas da face.

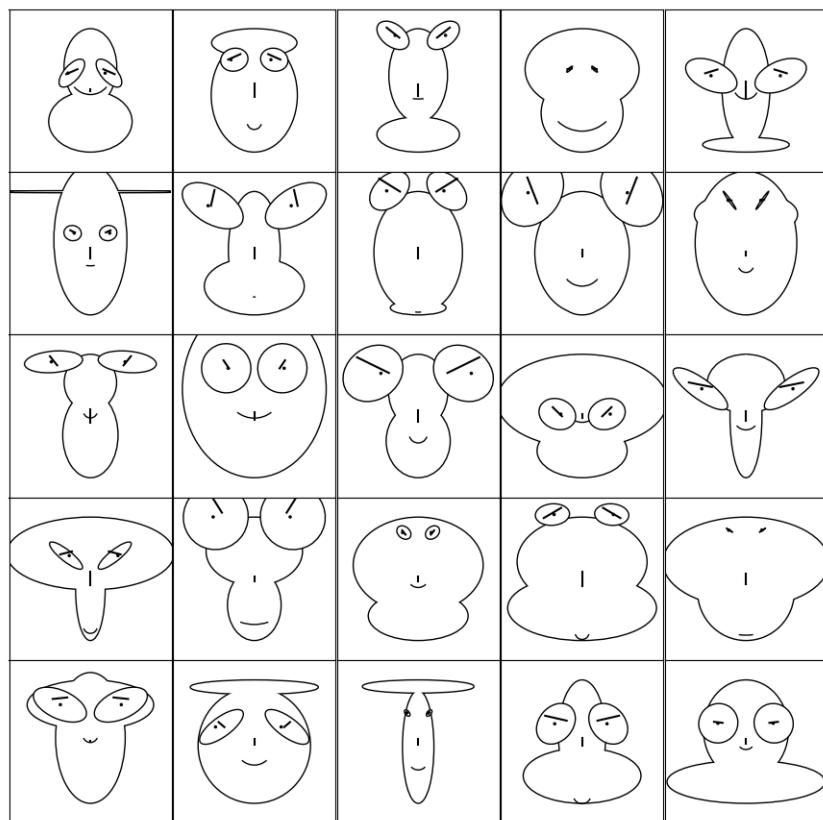


Figura 4.3: Diferentes faces de Chernoff para diferentes variáveis

**Proposição 4.2.1** *Faces de Chernoff foram criadas por Herman Chernoff em 1973, como uma forma de representar dados multivariados de forma a ser discernido facilmente por um observador humano. As faces constituem em linhas desenhadas em duas dimensões que contém uma série de estruturas faciais.* - Morris et al. [2000]

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, pois a análise deve ser realizada sobre as proporções gerais dos dados. 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 (4.4) demonstra esta situação graficamente. Notamos que 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.

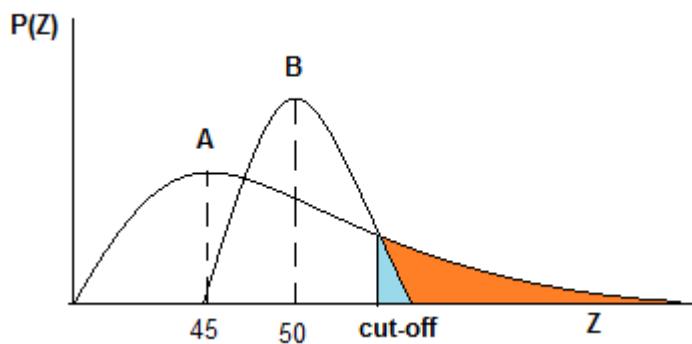


Figura 4.4: Exemplo de duas distribuições A e B relatando um depósito mais rico A com média menor que B. Área azul mostrando a contribuição da distribuição B acima do cut-off e área laranja mostrando a contribuição de A acima do cut-off

#### 4.2.1 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.

**Proposição 4.2.2** *Se lançarmos um dado de seis lados centenas de vezes, e anotarmos o valor realizado em cada jogada, teremos uma tabela com cada número e sua possível frequência. É esperado que a média deste conjunto de dados seja 3.5, pois a frequência entre os números obtidos nos lançamentos será aproximadamente parecida  $(6 + 1)/2$ . Este valor apesar não é real, pois não podemos obter metades de uma face de um dado, mas representa o centro de dispersão destes valores.*

As medidas de tendência central mais comuns são a média aritmética, a moda,

a média ponderada e a mediana.

### Média aritmética

A média aritmética pode ser descrita segundo a equação (4.2) em que  $x$  são os valores das amostras e  $n$  o número de amostras. Se a média aritmética for calculada a partir de uma população finita de todos os seus elementos a média  $\bar{x}$  é equivalente ao valor esperado da variável  $\mu = E(X)$ .

$$\bar{x} = \frac{1}{n} \sum_{i=0}^n x_i \quad (4.1)$$

Muitas vezes é necessário calcular a média aritmética de um agrupamento de dados a partir de um histograma, por exemplo. Neste caso podemos calcular a média aritmética como

$$\bar{x} = \frac{1}{n} (f_1 c_1 + f_2 c_2 + \dots + f_n c_n) = \frac{1}{n} \sum_{i=0}^n f_i c_i \quad (4.2)$$

Em que  $f_i$  é a frequência de cada classe  $c_i$ .

### Moda

Para variáveis inteiras, a informação mais importante é a frequência de cada valor da variável. Neste caso uma das informações importantes de tendência central é a moda, como o valor com maior frequência nos dados. No caso de variáveis reais contínuas, frequências são desprovidas de significado, sendo impossível calcular seus valor nas amostras, apenas por classes.

**Definição 4.2.2 — Moda.** A moda  $M_0$  de uma amostra é a observação com maior frequência nos dados

A moda nem sempre é um valor fixo, pois diferentes valores ou classes podem possuir mesma frequência. Quando um histograma apresenta dois picos, este também é chamado de **bimodal**. Quando apenas um é apresentado, chamamos o histograma de **unimodal**

### Média ponderada

A média ponderada considera que cada valor pode possuir uma importância diferenciada, e a ele é associado um valor chamado **peso**. A equação (4.3) demonstra o

valor de uma média ponderada

$$\bar{x} = \frac{\sum_{i=0}^n p_i x_i}{\sum_{i=0}^n p_i} \quad (4.3)$$

Em que  $p_i$  corresponde o peso de cada um dos valores para as  $n$  variáveis possíveis. A relação de cada peso pela soma total destes pesos também é chamado de ponderador e pode ser representado pela equação (4.4)

$$\lambda_i = \frac{p_i}{\sum_{i=1}^n p_i} \forall i \quad (4.4)$$

A média ponderada pode ser reescrita em termos de seus ponderadores de acordo com a equação (4.5)

$$\bar{x} = \sum_{i=0}^n \lambda_i x_i \quad (4.5)$$

### Mediana

A mediana é uma representação do valor associado a aproximadamente 50% da frequência total dos dados.

**Definição 4.2.3 — Mediana.** Se o número de elementos ( $n$ ) for ímpar, a mediana é igual a  $\frac{n+1}{2}$  elemento. Se o número de elementos for par, então a mediana é igual a média do  $\frac{n}{2}$  elemento e o  $\frac{n}{2} + 1$  elemento

**Proposição 4.2.3** Suponha que a amostra consiste em 10 observações: 6, 3, 4, 7, 4, 6, 7, 6, 5, 3, nós teremos um número de elementos  $n = 10$ , sendo este valor par. Ordenando o conjunto de dados teremos 3, 3, 4, 4, 5, 6, 6, 6, 7, 7. Então a mediana é igual a média entre o 5º e o 6º elemento, correspondendo ao valor de  $(5 + 6)/2 = 5,5$ .

## 4.2.2 Medidas de posição

As medidas de posição são aquelas tomadas em relação a outras, ou seja em seu contexto geral com outros valores. Entre elas as mais comuns são os **percentis**, **quartis**, e **decis**

### Percentis ou quantil

Uma das formas de se avaliar a posição dos dados é quanto a sua frequência. Um percentil ou quantil representa o valor correspondente a uma proporção total dos

dados.

**Definição 4.2.4 — Percentil ou quantil.** Um percentil ou quantil  $c_p$  de uma amostra corresponde ao valor imediatamente superior ou igual a  $100xp\%$  e imediatamente inferior a  $100x(1 - p\%)$  dos dados

**Proposição 4.2.4** Suponha que a amostra a seguinte amostra:  $\{6, 3, 4, 7, 4, 6, 7, 6, 5, 3, 4, 2\}$  nós teremos um número de elementos  $n = 10$ . Ordenando o conjunto de dados teremos  $\{2, 3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7\}$ . logo as proporções dos dados serão  $\{2 : 8\%, 3 : 17\%, 4 : 25\%, 5 : 8\%, 6 : 25\%, 7 : 17\%\}$ . As proporções acumuladas serão equivalentes a  $\{2 : 8\%, 3 : 25\%, 4 : 50\%, 5 : 58\%, 6 : 83\%, 7 : 100\%\}$ . Então o percentil de 67% será o valor imediatamente superior a 58% e inferior a 83%. Utilizando uma interpolação linear temos que  $(67\% - 58\%) * (6 - 5) / (83\% - 58\%) + 5 = 5.36$ .

### Quartis

O **quartil** são medidas de posição que correspondem a 4 posicionamentos especiais dentro do conjunto de dados. O primeiro quartil representa o **o percentil de 25%**, o segundo quartil representa **o percentil de 50% ou a mediana**, e o terceiro quartil representa **o percentil de 75% dos dados**.

**Proposição 4.2.5** . 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.

## 4.2.3 Medidas de dispersão

Outras medidas importantes são as de dispersão. Entre as mais comuns podemos citar a **variância**, o **desvio padrão** e a **amplitude** dos dados.

### Amplitude

A forma mais simples de se medir a dispersão dos dados é considerar sua amplitude. A maior vantagem em se definir a amplitude é sua simplicidade de cálculo, porém esta estatística é muito afetada por valores extremos

**Definição 4.2.5 — Amplitude.** Corresponde a diferença do valor máximo obtido nos dados  $x_{max}$  com o valor mínimo  $x_{min}$ .

### Intervalo Interquartil

Uma forma de se avaliar uma medida de dispersão menos afetada pelos valores extremos é o intervalo interquartil

**Definição 4.2.6 — Intervalo interquartil.** Corresponde a diferença do valor do terceiro quartil ( $Q_{75}$ ) com o valor do primeiro quartil ( $Q_{25}$ ).

### Variância

A variância pode ser descrita pela equação (4.7)

$$s^2 = \frac{\sum_{i=0}^n (x_i - \bar{x})^2}{n - 1} \quad (4.6)$$

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 ( $s = \sqrt{s^2}$ ).

Em alguns casos também é possível calcular a variância para classes e não para valores, assim como a média aritmética. Neste caso podemos calcular a variância a partir de

$$s^2 = \frac{\sum_{i=0}^n f_i (c_i - \bar{c})^2}{n - 1} \quad (4.7)$$

Em que  $c$  corresponde ao valor da classe e  $f_i$  o valor da frequência associada aquela classe.

#### 4.2.4 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 (4.5) 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.

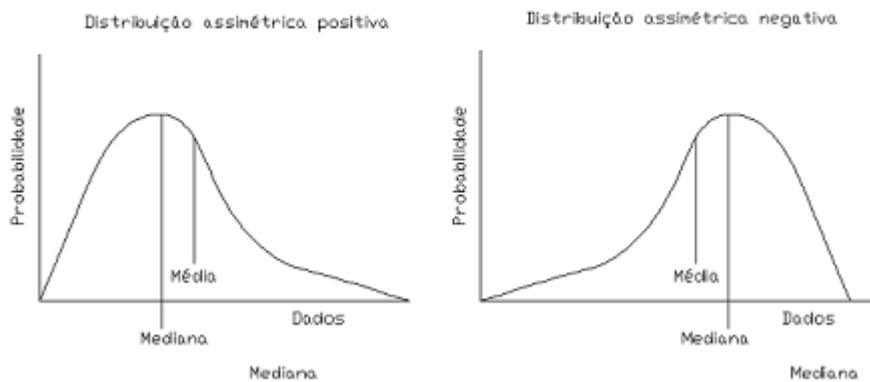


Figura 4.5: Assimetria de uma distribuição de dados a) Assimetria positiva b) assimetria negativa

Uma das medidas de assimetria mais comuns é o coeficiente de Pearson que pode ser expresso pela equação (4.8)

#### Coefficiente de assimetria de Pearson

$$S_p = 3(\bar{x} - M_e) / s \quad (4.8)$$

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

**Proposição 4.2.6** Imagine uma variável com valor de média  $\bar{x} = 198.89$ , valor de mediana  $M_e = 128.15$  e valor de desvio padrão igual a  $s = 180.56$ . O valor do coeficiente de assimetria será igual a  $3(192.89 - 128.30)/180.56 = 1.07$ , demonstrando assimetria nos dados.

Distribuições com característica de assimetria positiva são muito comuns na avaliação de depósitos minerais, principalmente no tratamento de commodities 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.

#### 4.2.5 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 bastariamos 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 (4.9)

$$CV = \frac{s}{\bar{x}} \quad (4.9)$$

Os coeficientes de variação são medidas importantes para a pesquisa mineral, porque são a primeira forma utilizada para classificar depósitos minerais segundo sua regularidade. O livro de [Maranhao, 1985] demonstra a classificação de depósitos minerais de acordo com o coeficiente de variação, tal como na tabela 4.1.

Tabela 4.1: Regularidade dos depósitos minerais de acordo com a classificação do coeficiente de variação

Regularidade	Coeficiente de variação	Exemplo
Regulares	$5\% < CV < 40\%$	Jazidas de ferro, manganês, níquel, cobalto
Irregulares	$40\% < CV < 100\%$	Jazidas de fluorita, barita, grafita, corídon
Muito irregulares	$100\% < CV < 150\%$	Jazidas de tungstênio em tactitos, ouro
Extremamente irregulares	$CV > 150\%$	Pegmatitos com berilo, tantalita, columbita

Valores de coeficiente de variação maiores representam geralmente um maior desafio para a aplicação de técnicas de geoestatística, pois geralmente apresentam alta variabilidade ou erraticidade dos dados.

#### 4.2.6 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  $(\bar{x}) \pm ks$ . 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 (4.10) como a desigualdade de Chebyshev.

$$P(|X - \mu| \geq k\sigma) \leq 1/k^2 \quad (4.10)$$

Em que  $X$  é o valor da variável aleatória,  $\mu$  é o valor da média da população,  $\sigma$  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

a variável aleatória. Apesar de não possuirmos os valores  $(\mu, \sigma)$  correspondentes aos parâmetros da população, podemos estimar os valores a partir das estatísticas das amostras. Se o número de amostras for grande o suficiente e as técnicas de amostragem bem selecionadas, podemos dizer que  $(\mu \sim \bar{x}, \sigma \sim s)$

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  $(50 \pm 20) \text{ ppm}$  em que 20 representaria dois desvios padrões de 10 ppm e 50 ppm seu valor médio.

### 4.3 Validação do banco de dados e 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. Análises realizadas com valores muito discrepantes pode acabar gerando resultados espúrios e inconsistentes com a realidade.

**Definição 4.3.1 — Outlier.** *Um outlier é considerado um valor ou observação caracterizado pela sua relação entre o restante de observações que fazem parte das amostras. O seu distanciamento em relação as observações é essencial para fazer sua caracterização. Estas observações também são chamadas de 'anormais', contaminantes, estranhas, extremas ou aberrantes - Figueira [1998]*

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. Machado [2012] indica que o surgimento de valores anômalos podem ocorrer por diversas formas, entre elas:

1. **Valores errôneos:** As possíveis causas são os erros de análise ou de digitação, troca de amostras, contaminações de amostras ou até mesmo fraude.
2. **Valores pertencentes a outra população:** Podem ocorrer devido à mistura de diferentes teores ou litologias, ou que possuem processo formacional em tempos geológicos distintos. A revisão dos domínios geológicos, neste caso é recomendado, de forma a tratar e estimar os dados separadamente.

**3. Valores pertencentes a mesma população:** Podem ocorrer eventos metagenéticos que favoreçam a concentração de uma propriedade em parte do depósito. Estes eventos estão relacionados também ao chamado **efeito pepita**, em que proporções erráticas podem aparecer apenas em locais distintos do depósito, em regiões pequenas.

A Tabela 4.2 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 4.2: Tabela de teores do minério de ferro

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

**Proposição 4.3.1** *Pode até mesmo parecer um clichê, mas a melhor forma de se analisar outliers é com bom senso. Devemos entender o problema, analisá-lo profundamente na hora de limparmos o banco de dados. Permitir que bancos de dados sejam transmitidos antes de uma boa verificação pode resultar no fracasso de uma análise destes dados.*

Diversas são as formas de identificação de valores outliers. Técnicas para valores em apenas uma variável são muito conhecidas, no entanto, deve-se entender que um valor anômalo depende de sua dimensão analisada. Uma amostra outlier considerando variáveis distintas pode não ser um valor um valor anômalo quando considerado um problema multivariado.

Uma das ferramentas mais comuns para identificação de valores anômalos é o gráfico 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 das retas formadas pela adição e subtração 1,5 vezes o intervalo interquartil dos valores máximo e míno dos dados representam outliers. O intervalo interquartil é também determinado como a diferença entre os valores do terceiro quartil e do primeiro quartil. A figura 4.6 demonstra o gráfico boxplot e suas dimensões.

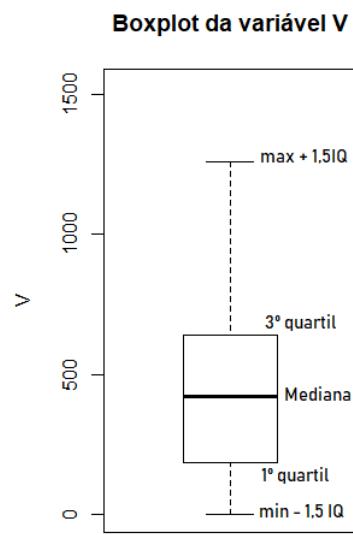


Figura 4.6: Representação de um gráfico de caixa dividida entre os intervalos das amostras

Os valores anômalos ou outliers são demonstrados na figura 4.8 como pontos circulados fora das barras que representam os limites de aceitação dos valores da amostra.

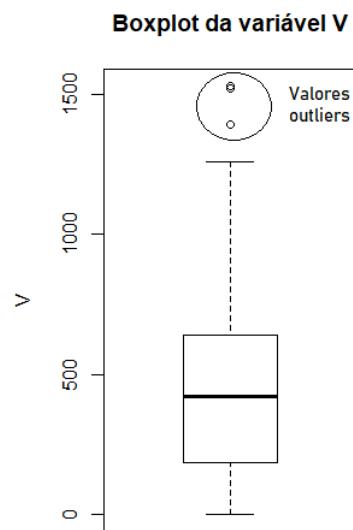


Figura 4.7: Representação dos valores outliers no gráfico boxplot - Pontos circulados em vermelho

Muito cuidado deve ser utilizado com esta ferramenta. Em alguns casos distribuições de dados assimétricas podem gerar no gráfico boxplot uma quantidade de valores anômalos absurdas. A melhor forma de lidar com valores outliers é o bom senso, ferramentas são úteis, mas não devem ser o critério determinante na maioria dos casos.

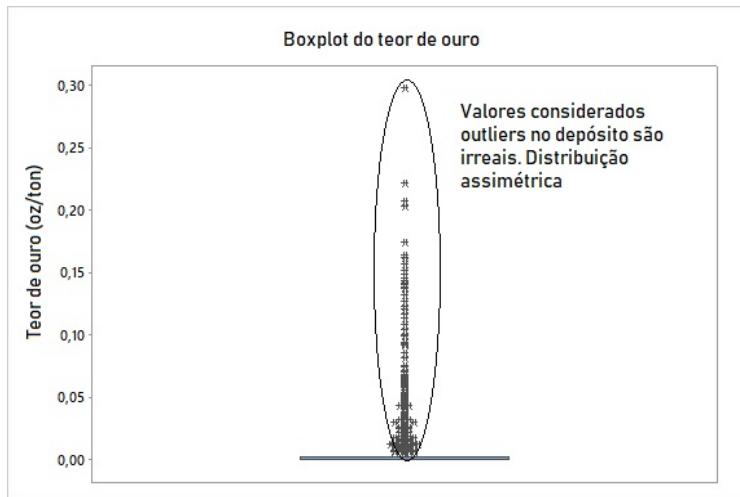


Figura 4.8: Valores outliers em uma distribuição assimétrica dos dados. Nota-se que grande parte da informação é considerada outliers. Neste caso é necessário bom senso para não se remover informações desnecessárias e prejudicar a análise de dados.

Após a identificação de valores anômalos é possível realizar o tratamento destes dados. É imprescindível entender que bancos de dados **nunca** devem ser alterados, apenas estatísticas. A alteração ou remoção de dados é considerada uma atitude imoral para analistas de dados.

- R** É importante entender que um banco de dados **nunca** deve ser alterado. Apenas as estatísticas são cabíveis a manutenção. A alteração de dados reais pode ser considerada um ato imoral, principalmente na mineração, onde o trabalho, segurança e condições de vidas de muitas pessoas estão em jogo.

Diversas alternativas podem ser utilizadas para o tratamento de valores outliers. Dentre elas podemos citar

1. **Truncamento:** Após identificar valores outliers é possível normalizar seus valores para os valores extremos (máximo ou mínimo), ao desconsiderá-los. O truncamento de dados na geoestatística deve ser feito de forma a evitar que os valores anômalos não alterem significativamente as estatísticas globais. Como regra de ouro considera-se que o truncamento deve ser feito sem que se altere mais do que 10% dos valor médio das amostras.

2. **Remoção:** Em alguns casos a remoção dos valores outliers pode ser feita. Se a proporção de dados removidos for alta, é possível alterar excessivamente as estatísticas, por isso muito cuidado deve ser feito ao considerar uma amostra como outlier.
3. **Reescalonamento:** Dependendo da distância relativa dos outliers com o contexto geral das amostras é possível realizar uma redução de suas distâncias até o valor máximo desconsiderando-os.

## 4.4 Descrição espacial das amostras

A geoestatística é uma ciência que prevê a utilização de informações no espaço, e para isso muitas vezes utilizamos informações de mapas. Mapas são representações visuais de uma região que são dotados de informações como **escala, legenda, título, orientação**.

Mapas de localização destas amostras são uma ferramenta gráfica muito importante para determinar o comportamento de variáveis no espaço. Mapas devem ser feitos de forma cuidadosa, representando escalas condizentes com o objeto de estudo e garantindo a melhor visualização possível das amostras.



*A qualidade desejada de um mapa varia de acordo com a investigação. Tipicamente a representação de um mapa deve ser limpa, sem valores altos ou baixos ambíguos, e mostrar os dados o menos distorcidos possível com um mínimo de artefatos computacionais - Gustavsson et al. [1997]*

Estas informações nos permitem identificar regiões consideradas mais ricas, regiões onde ocorrem agrupamentos característicos dos dados, e o layout das malhas de amostragem. A figura 4.9 demonstra um depósito polimetálico de Jura. O atributo é o tipo de rocha de um dado período geológico.

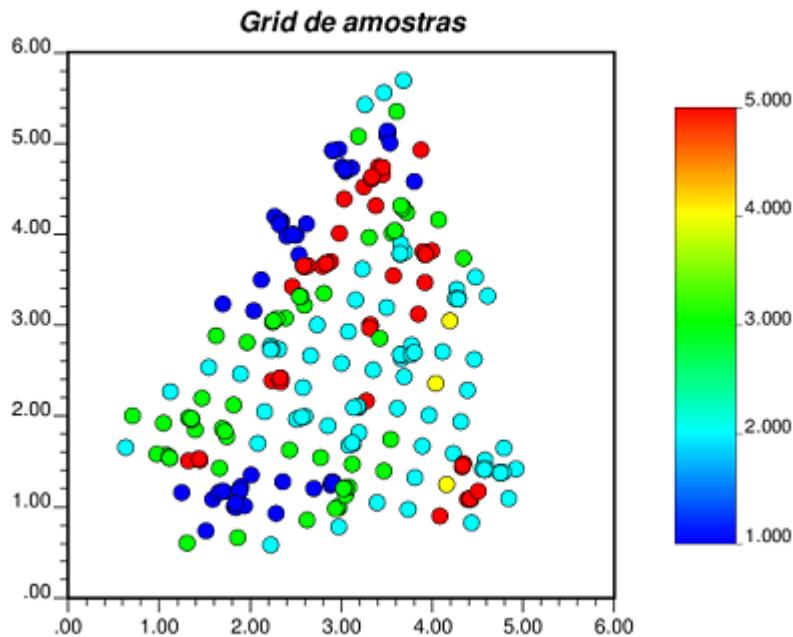


Figura 4.9: Disposição das amostras no espaço. Cores diferenciadas mostrando tipos de rocha em períodos geológicos diferentes

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 exploração cujo o interesse seria o litotipo 1, provavelmente começariámos a retirar o material de oeste para leste para reduzir o fluxo de caixa do empreendimento.

A 4.10 demonstra a propriedade de teor de Cádmio obtida nestas amostras no depósito. Podemos verificar sua distribuição segundo esta disposição deltaica.

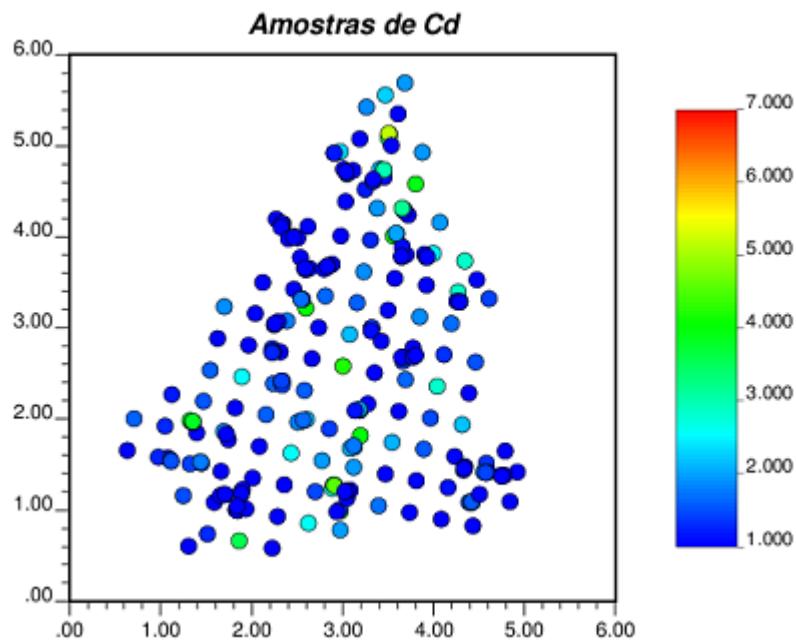


Figura 4.10: Disposição do Cd

As informações disponíveis nestes mapas nos permite associar as informações entre as variáveis do tipo litológico e o teor de Cádmio. 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. Esta análise visual nos permite entender o comportamento de certas variáveis e sua disposição no espaço, buscando explicações para os valores destas propriedades. Além das informações obtidas em mapa também podemos visualizar amostras e propriedades em um espaço tridimensional. A figura 4.11 demonstra a disposição de amostras e a visualização do comportamento de uma propriedade binária no espaço.

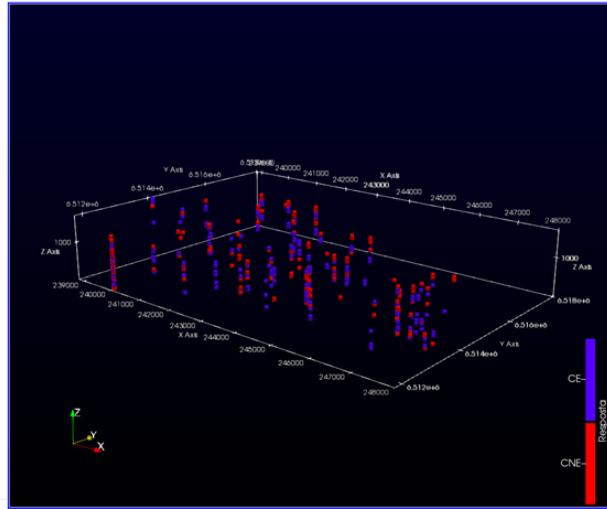


Figura 4.11: Informações de amostras obtidas em três dimensões.

## 4.5 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.

**Definição 4.5.1 — histograma.** *Um histograma é uma ferramenta gráfica, representada por um gráfico de barras que condiciona os valores de uma variável com suas frequências.*

Esta ferramenta é essencial principalmente em três condições:

1. **Classificação:** Quando possuímos classes distintas o histograma apresenta diretamente a proporção de cada classe considerada
2. **Contagem:** Quando a variável constitui em valores inteiros, cada valor desta variável pode ser diretamente associada a sua frequência.
3. **Contínuo:** Quando os valores são reais, podemos atribuir intervalos de classe (ou em inglês *bins*) aos quais estes valores estão inseridos. Dependendo do número de intervalos de classe e seu tamanho o histograma pode apresentar diferentes formas.

Uma das proposições utilizadas para o cálculo do número ótimo de intervalos de classes é pela fórmula de Sturges 4.11

$$\hat{h} = \frac{\text{amplitude dos dados}}{1 + \log(n)} \quad (4.11)$$

Em que a amplitude dos dados é relacionado a diferença do máximo e do mínimo das amostras e  $n$  é o número de amostras. A figura 4.12 representa um histograma da variável Cádmio do depósito de Jura. Podemos notar como a distribuição dos dados se comporta nesta variável, como aspectos de simetria, valores médios, e inclusive possíveis valores outliers, quando as barras de frequência são pequenas e distanciadas da maioria.

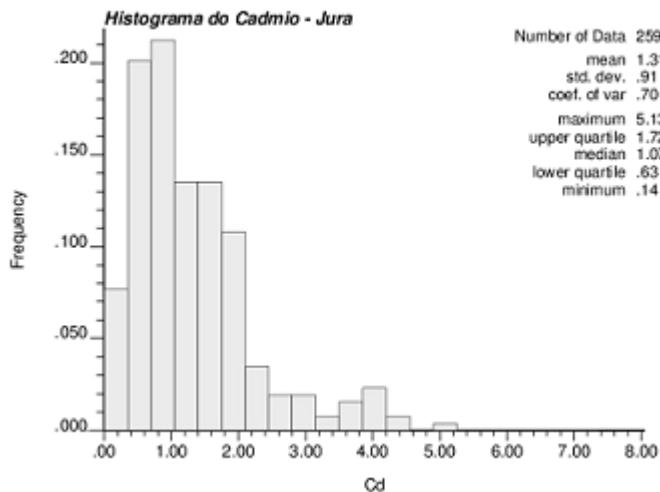


Figura 4.12: Histograma do Cd

A observação de uma frequência de uma classe é diretamente relacionada ao tamanho desta. Na figura 4.13 podemos ver que a classe de teores de 0,04 a 0,75g/ton ocupa uma proporção de 20 % dos dados.

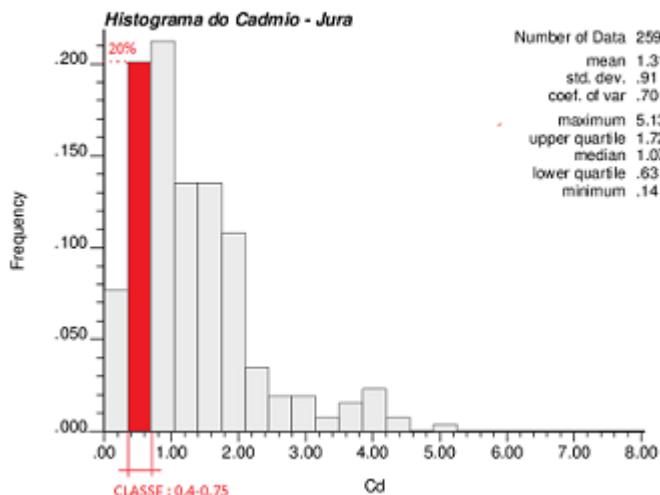


Figura 4.13: Histograma do Cd - Classe marcada

A construção de histogramas envolve sempre a criação de intervalos de classe de mesmo tamanho. Alterar apenas o tamanho de uma classe em detrimento de outras pode ser considerado um uso abusivo das estatísticas, enviesando a percepção de outras pessoas sobre as verdadeiras frequências dos dados.

**R**

Não é correto alterar o tamanho de apenas um intervalo de classe em detrimento dos outros. Esta prática é mal vista, e pode ser intuitivamente criada para gerar vies na percepção dos leitores quanto as frequências de determinados valores.

Assim como em outras estatísticas, a utilização dos histogramas favorece o entendimento global dos dados, mas prejudica no entendimento apurado da variável. A escolha do tamanho do intervalo é uma variável importante para a observação desta estatística gráfica. Valores de classe com tamanho muito grande apresentaram frequências maiores, mas perderão a forma natural dos dados. Valores de classe com tamanho muito pequeno apresentarão baixa frequência e se tornarão mais achados, dificultando a visualização das proporções da variável.

**Proposição 4.5.1** *A escolha do tamanho do intervalo de classe é fundamental para verificar a forma do histograma e sua representação real. Valores de tamanho muito pequenos ou grandes podem gerar gráficos pouco intuitivos, escondendo a real simetria, valores médios e dispersão dos dados. É importante que um histograma caracterize visualmente os dados de forma a representar as estatísticas numéricas a serem calculadas.*

Outra forma de representar um histograma é na sua forma acumulada. Neste caso cada valor das frequências de uma variável são aumentadas em ordem crescente, do menor valor das amostras até o maior valor. A figura 4.14 é uma demonstração do gráfico acumulado.

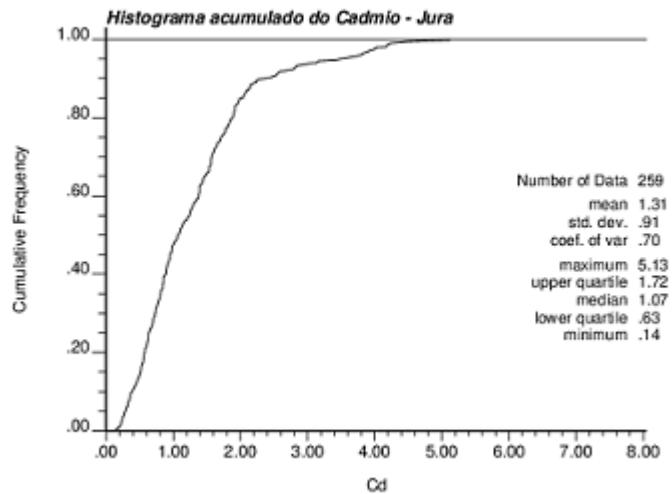


Figura 4.14: Histograma do Cd acumulado

A figura 4.15 demonstra a leitura do gráfico acumulado. Podemos notar por este gráfico que 60 por cento dos valores estão abaixo do teor de 1,5g/tonelada.

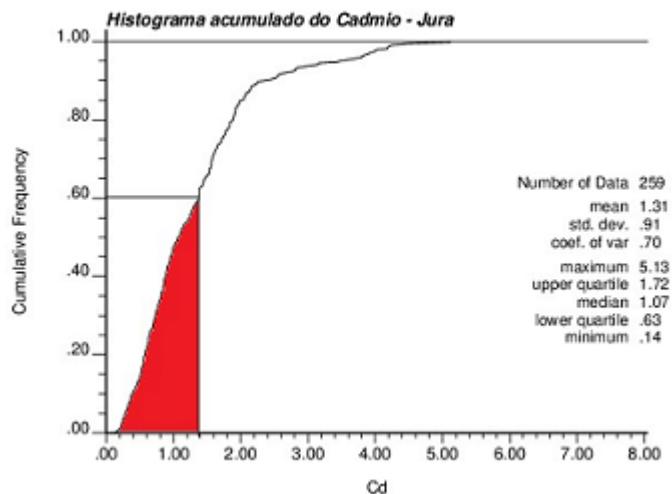


Figura 4.15: Histograma do Cd acumulado - Leitura

O formato dos histogramas pode adicionar importantes informações sobre a distribuição dos dados, como por exemplo a assimetria. Na figura 4.16 podemos observar dois histogramas de depósitos minerais diferentes, um simétrico de Ferro em A) e um de alumínio assimétrico em B). Quando consideramos as técnicas clássicas de avaliação de depósitos a assimetria dos dados pode dificultar os métodos convencionais, o que torna depósitos de alta assimetria mais difíceis de reproduzirem estimativas condizentes.

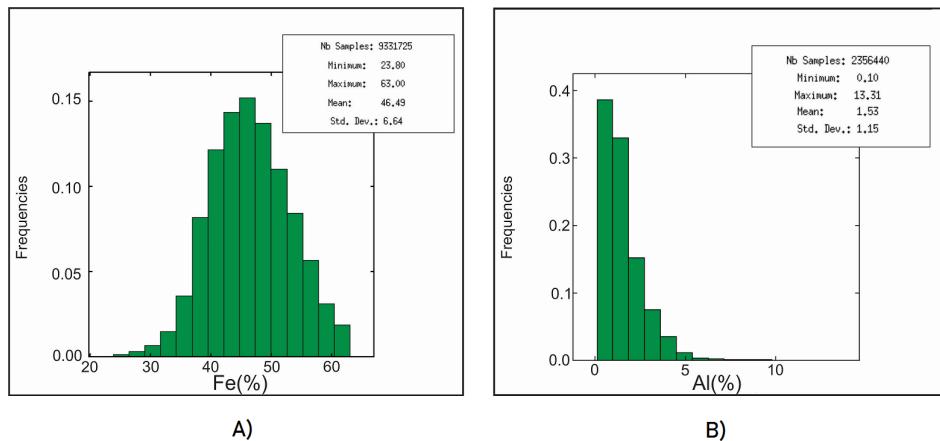


Figura 4.16: Simetria para diferentes histogramas - a) histograma simétrico, b) histograma assimétrico

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. Distribuições de frequências centradas podem ter como candidato um modelo de ajuste gaussiano, por exemplo. Distribuições assimétricas podem se encaixar, por exemplo, em um modelo lognormal.

## 4.6 Inferência Estatística

Após analisados os dados amostrais podemos utilizar funções para modelar populações dos dados. Na maioria dos casos não precisamos conhecer *a priori* as distribuições da população, mas em alguns casos como na geoestatística não-linear, conhecer uma distribuição teórica de probabilidades pode facilitar estudos para entender problemas mais complexos

**Definição 4.6.1 — Inferência estatística.** *Inferência estatística é o método pelo qual deduzimos informações da população dos dados com base em informações das amostras*

,

### 4.6.1 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  $x$  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 de rocha a quantidade de fraturas seja pequena, enquanto para tamanhos grandes de rocha essa densidade aumente.

A função de distribuição de Poisson pode ser escrita segundo a equação (4.12)

$$P(X = x) = \frac{\exp^{-\lambda} \delta^x}{x!} \quad (4.12)$$

Em que  $x$  é uma realização da variável aleatória  $X$ ,  $P(X = x)$  é a probabilidade associada àquele evento e  $\lambda = E(X)$  sendo o parâmetro da função. Na maioria dos casos aproximamos  $E(X) \sim \bar{x}$ . 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 :

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

Então:

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

■

### Distribuição Gaussiana

Esta talvez seja uma das funções de densidade de probabilidade mais populares e representa um grande papel na geoestatística. As equações de estimativa lineares que serão apresentadas neste livro são também analogamente chamadas de **equações normais**. Isto se deve pelo fato de que os resultados obtidos em variáveis gaussianas são os mais precisos possíveis dentro de todas outras distribuições na geoestatística. Quanto mais próximo for a distribuição das amostras de uma distribuição gaussiana, melhores serão os resultados de uma estimativa geoestatística.

**Proposição 4.6.1** Consideremos uma variável  $Z$ , gaussiana e estacionária (em prática a variável que pode ser aproximada de um histograma por uma gaussiana), com média  $m$  e variância  $\sigma^2_Z$ , a hipótese de permanência da normalidade indica que uma variável  $Y$  estimada segue uma distribuição de mesma forma e média  $m = E(Z) = E(Y)$  e variância  $\sigma^2_Z \neq \sigma^2_Y$  -Journel and Huijbregts [1978]

O formato de uma distribuição gaussiana é tipicamente na forma de um sino (*bell shape*), centrado em um valor médio e com uma variância característica. A figura 4.17 demonstra uma distribuição gaussiana típica com média igual a 5 e variância igual a 2.

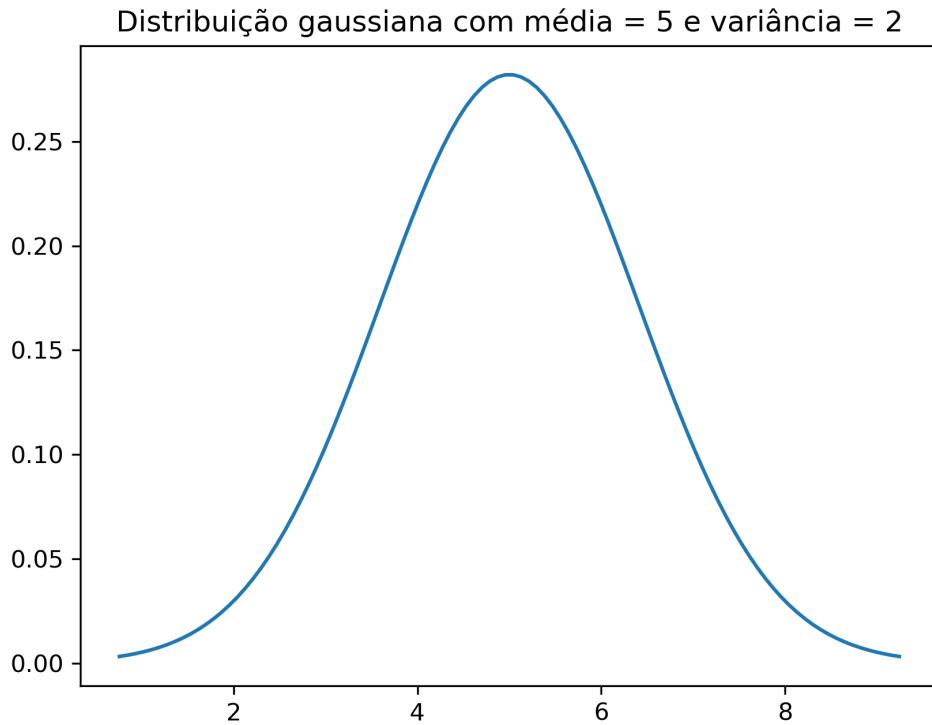


Figura 4.17: Forma de uma distribuição gaussiana com média 5 e variância 2

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 (4.13)

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4.13)$$

Em que  $\sigma^2$  é a variância da distribuição aleatória e  $\mu$  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 (4.14)

$$P(X = x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} \quad (4.14)$$

Uma variável aleatória pode ser padronizada segundo a relação (4.20)

$$X_p = (X - \mu)/\sigma \quad (4.15)$$

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. A integral da função gaussiana não possui uma antiderivada para utilizarmos explicitamente, por isso o truque utilizado é provar que o quadrado da integral da gaussiana é equivalente a  $2\pi$ . Logo temos:

*Demonstração.* Prova do somatório de uma função gaussiana ser igual a 1

$$\begin{aligned} Int^2 &= \left( \int_{-\infty}^{\infty} e^{-\frac{(x)^2}{2}} dx \right)^2 = 4 \int_0^{\infty} e^{-\frac{(t)^2}{2}} dt \int_0^{\infty} e^{-\frac{(u)^2}{2}} du \\ &4 \int_0^{\infty} \int_0^{\infty} e^{-\frac{(t^2+u^2)}{2}} dt du \end{aligned}$$

Alterando para coordenadas polares

$$\begin{aligned} &4 \int_0^{\infty} \int_0^{\pi/2} r e^{-\frac{r^2}{2}} dr d\theta \\ &2\pi \int_0^{\infty} r e^{-\frac{r^2}{2}} dr \\ &2\pi \end{aligned}$$

Logo se:  $Int^2 = 2\pi$

$$Int = \sqrt{2\pi}$$

portanto :

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1$$

■

### 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 seu logaritmo é normalmente distribuído. A equação (4.16) demonstra a função de densidade de probabilidade para a distribuição lognormal.

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp^{\frac{(-\log(x))^2}{2}} \quad (4.16)$$

O Valor esperado da distribuição pode ser demonstrado segundo a equação (4.17)

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \quad (4.17)$$

A figura 4.18 apresenta a forma assimétrica da distribuição lognormal, para um distribuição com média 5 e variância igual a 2.

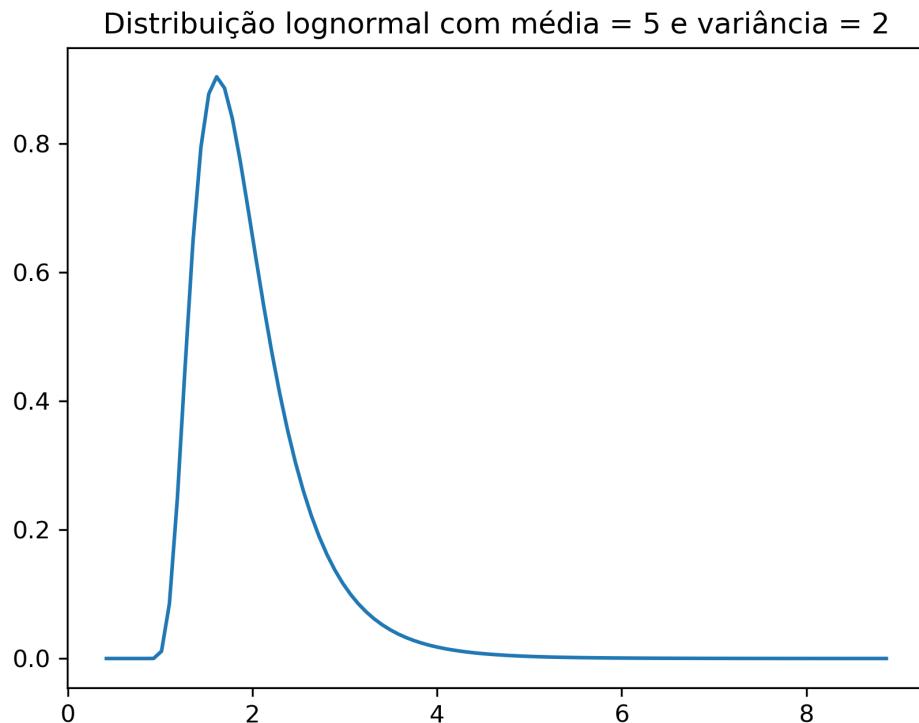


Figura 4.18: Forma de uma distribuição lognormal com média 5 e variância 2

### Estimando a média da população

O processo de inferência estatística resume-se em determinar características da população a partir de dados amostrais. Podemos estimar o valor real da média da função aleatória  $Z(x)$  a partir do estimador  $\hat{Z}(x)$  a partir da média aritmética  $\sum_{i=1}^n Z(x_i)/n$  em que  $n$  constitui um número grande de variáveis aleatórias em diferentes suportes  $i$ . A equação (4.18) apresenta este processo.

$$E(\hat{Z}(x)) = E\left(\sum_{i=1}^n Z(x_i)\right)/n = \left(\sum_{i=1}^n E(Z(x_i))\right)/n = \left(\sum_{i=1}^n m\right)/n = m \quad (4.18)$$

Sobre a hipótese de estacionaridade da média, sabemos que a média das variáveis aleatórias é igual a média da função aleatória. 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 (4.19)

$$Var(m) = Var \left( \sum_{i=1}^n Z(x_i)/n \right) = \sum_{i=1}^n 1/n^2 Var(Z(x_i)) = \sigma^2/n \quad (4.19)$$

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. Quanto maior a quantidade de amostras utilizadas em uma estimativa, menores serão os erros associados a esta estimativa média local. A figura (4.19) demonstra como o valor médio tende a cada vez se aproximar mais da média das amostras com o aumento do número de amostras.

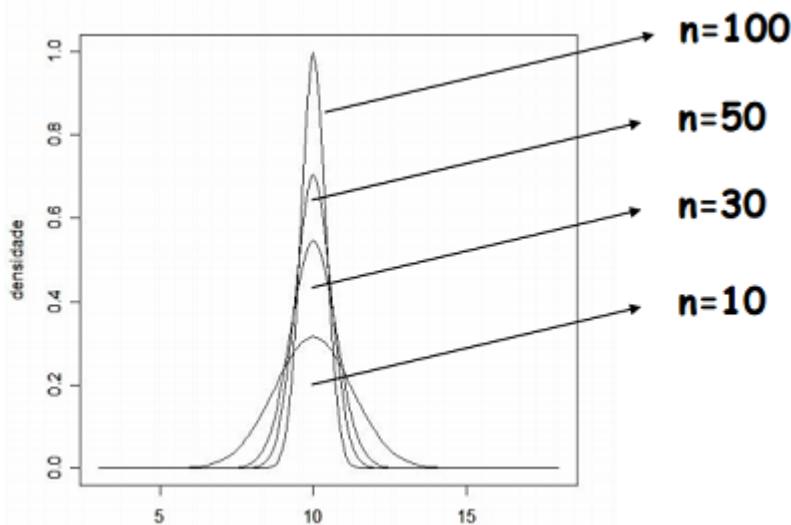


Figura 4.19: Figura demonstrando o efeito de suporte para um número crescente de amostras. O aumento do número de amostras tende a concentrar a função de densidade de probabilidade entorno do valor médio

## 4.7 Distribuição t-Student

Para determinarmos a distribuição gaussiana geralmente assumimos o conhecimento a respeito da variância da população. Se considerarmos a **distribuição de valores**

médios, sabemos que se  $Z_{x1}, Z_{x2}, \dots, Z_{xn}$  são amostras normalmente distribuídas normalmente  $\phi(m, \sigma^2)$ , então a quantidade

$$Z_p = \frac{(\bar{Z(x)} - \mu)}{\sigma/\sqrt{n}} \quad (4.20)$$

É distribuída com variável aleatória  $\phi(0, 1)$ . A distribuição dos valores médios  $(\bar{Z(x)} - \mu)/(\sigma/\sqrt{n})$  segue a distribuição chamada de t-Student, com  $n - 1$  graus de liberdade. Quando o número de amostras tende a crescer, aproximadamente de 30, a distribuição t-Student converge para a distribuição normal padrão  $\phi(0, 1)$ . Por isso dizemos que para estudos estatísticos iniciais, precisamos de pelo menos 30 amostras para se ter uma melhor compreensão da média.

## 4.8 Dimensionamento de malhas regulares

Em campanhas de prospecção preliminares é rotineiro utilizar técnicas estatísticas convencionais para estimar o tamanho e posicionamento de malhas de amostragem. No estágio inicial é necessário cobrir uma certa área de forma a verificar suas potencialidades. A medida que os estudos avançam, as amostragens tendem a aumentar e se tornarem mais densas, e estudos geoestatísticos mais avançados são realizados. A área de influencia de uma perfuração pode ser calculada pela equação (4.22)

$$A_0 = \frac{A}{n} \quad (4.21)$$

Os estudos iniciais são fortemente afetados pela regularidade do depósito mineral. Depósitos erráticos como veios de ouro tendem a necessitar de malhas mais adensadas que depósitos regulares como os de carvão mineral.

**R** O principal fator que controla a densidade da malha de perfuração é a regularidade do depósito e, por isso, a malha tem de ser cada vez mais densa, à medida que trabalham depósitos onde a variabilidade na forma ou qualidade (teor e conteúdo) é maior - Maranhao [1985]

Para encontrarmos o número mínimo de amostras segundo o erro esperado para amostragem, utilizamos a equação

$$N = \frac{(t.CV)^2}{E^2} \quad (4.22)$$

Em que  $t$  é o valor da variável t-Student para um nível de confiabilidade,  $CV$  é o valor do coeficiente de variação do depósito mineral e  $E$  é o valor do erro aceitável para a estimativa.

**Proposição 4.8.1** *Considere um depósito mineral com coeficiente de variação igual a 51,98%, um valor de confiabilidade para a média de 95% ( $t$ -student = 2.20, para 12 amostras), e um erro aceitável para uma medida de no máximo 20%. A área pesquisada é igual a  $70.000m^2$ , e realizaremos 12 amostras. Logo o erro que cometemos é  $E = \sqrt{\frac{(t.CV)^2}{N}} = \sqrt{\frac{(2.20 \cdot 51.98)^2}{12}} = 32.7\%$*

## 4.9 Exercícios

**Exercícios 4.1** 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% ■

**Exercícios 4.2** 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.:  $\bar{x} = 5.7$ ,  $s^2 = 5.06$ ,  $s = 2.25$  ■

**Exercícios 4.3** 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 ■



## 5. Estatística bivariada

*Como em outras artes, a ciência da dedução e análise é uma que não pode ser adquirida por um longo e paciente estudo, nem é a vida longa o suficiente para permitir qualquer mortal se ater a mais alta perfeição nela.*

*Sherlock Holmes em 'Um estudo em Vermelho'*

### 5.1 Introdução

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. Inicialmente apresentamos as **ferramentas gráficas** mais utilizadas e depois algumas **estatísticas pontuais** utilizadas.

## 5.2 Probabilidade condicional e Esperança condicional

### 5.2.1 Probabilidades condicionais e conjuntas

Probabilidades não são nada além de métricas de conjuntos, proporções de acordo um espaço amostral ( $\Omega$ ). Estas proporções podem tomar diferentes características quando analisamos não apenas um conjunto individual, mas a interação entre eles. Muitas vezes não estamos interessados em determinar as probabilidades ou frequências individuais de uma variável aleatória. É interessante, por exemplo, determinar combinações entre variáveis e suas possíveis relações. E se desejarmos saber qual é a frequência de um minério e que seu conteúdo tenha um determinado valor de impureza? Se denotarmos  $X$  como o evento de ser minério, e  $Y$ , a variável que denota seu limite de impurezas, podemos denotar a probabilidade de  $P(X, Y)$  como sendo a probabilidade de que "*Um material seja minério e apresente impurezas acima do limite desejado*". A forma mais simples de se entender probabilidades é de acordo com um diagrama de Venn, como na figura 5.1

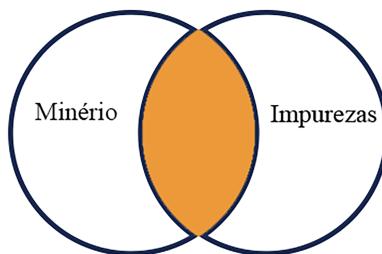


Figura 5.1: Demonstração de eventos disjuntos entre a variável minério e impureza a partir de um diagrama de Venn. Nota-se a área laranja como sendo a interseção dos eventos representado pela probabilidade  $P(X, Y)$

Observe a tabela 5.1. 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  $P(\text{Minério}) \cap P(\text{Impureza} \geq 0,005) = 2/5 = 40\%$

Tabela 5.1: Tabela da relação entre um dado minério e uma impureza

Minério	Impureza	$\text{Minério} \cap (\text{Impureza} \geq 0,005)$
Sim	0,005	1
Não	0,007	0
Não	0,008	0
Sim	0,006	1
Sim	0,003	0

Em alguns casos também é importante determinar a conjunção entre os eventos, ou a probabilidade de  $P(X \vee Y)$ . Neste caso queremos determinar "*Um material seja minério ou apresente impurezas acima do limite desejado*". Note que a conjunção 'ou' é um conectivo lógico muitas vezes dispare do seu uso corriqueiro no português. Ser um ou outro na verdade não é uma escolha entre um dos elementos, mas uma soma dos eventos. A representação da conjunção pode ser vista na figura 5.2.

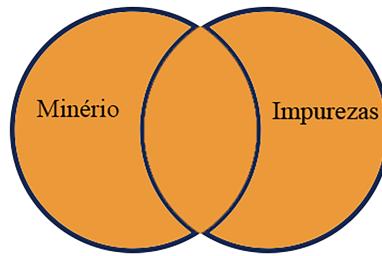


Figura 5.2: Demonstração da conjunção de eventos entre a variável minério e impureza a partir de um diagrama de Venn. Nota-se a área laranja como sendo a interseção dos eventos representado pela probabilidade  $P(X \vee Y)$

Estas relações lógicas envolvem o conhecimento entre os eventos independentemente. Conhecer  $P(X, Y)$  é exatamente o mesmo que conhecer  $P(B, A)$ . Em alguns casos devemos entender o conceito de dependência na estatística, expresso pela probabilidade condicional. Neste caso queremos saber "*dado que um material apresente impurezas, qual é sua probabilidade de ser minério*". Esta é uma afirmação muito diferente da obtida nos outros casos, pois para sabermos se algo é minério, precisamos saber antes se ele contém impurezas. A probabilidade condicional  $P(X|Y)$  pode ser demonstrada pela figura 5.3 como a relação da área laranja pela área hachurada.

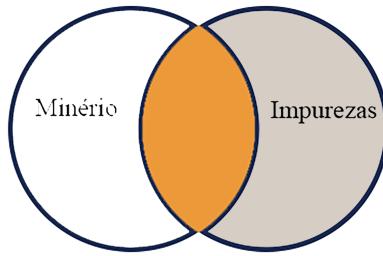


Figura 5.3: Demonstração da probabilidade condicional entre a variável minério e impureza a partir de um diagrama de Venn. Nota-se a área laranja como sendo a interseção dos eventos disjuntos. A probabilidade condicional é a relação entre a área laranja pela área hachurada.

Esta relação também é chamada de teorema de Bayes, e envolve a equação (5.1)

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \quad (5.1)$$

**Proposição 5.2.1** *As probabilidades condicionais expressam um importante conceito na geoestatística, a dependência entre variáveis aleatórias. Quando estudamos fenômenos espaciais, os valores obtidos em um suporte específico  $x_1$  são muito mais dependentes de  $x_2$  do que  $x_3$ , se a distância de  $\{x_1, x_2\}$  for menor que a distância de  $\{x_1, x_3\}$*

### 5.2.2 Esperança condicional

A partir da definição de probabilidade condicional também é possível determinar a esperança condicional. Ela nada mais é que o valor médio obtido de uma variável  $Y$  dado que a variável  $X$  assuma um valor específico  $x$ . Por exemplo, podemos determinar qual é a probabilidade do material ser contaminado  $Y$ , dado que a presença de um litotipo  $X$  seja  $x = \{\text{itabirito}\}$ , de acordo com a equação (5.2).

$$E(Y|X = x) = \sum_{y \in Y} y P(Y = y|X = x) \quad (5.2)$$

Considere que  $Y$  seja uma variável binária tal que assuma o valor 0 para o elemento contaminado, e valor 1 para não contaminado. A variável  $X$  pode assumir os valores de itabirito e calcário no problema. Analisando a tabela 5.2 podemos determinar as probabilidades de acordo com os respectivos valores apresentados.

Tabela 5.2: Tabela da relação entre um minério contaminado e litotipo

Y	X	$(Y=0   X=\text{itabirito})$	$(Y=1   X=\text{itabirito})$
contaminado	itabirito	1	0
descontaminado	calcário	0	0
descontaminado	calcário	0	0
contaminado	itabirito	1	0
descontaminado	itabirito	0	1
$P(Y=y   X=x)$		2/3	1/3

O valor esperado condicional pode ser obtido a partir da tabela pode ser calculado por (5.3)

$$E(Y|X = \text{itabirito}) = 2/3.0 + 1/3.1 = 1/3 \quad (5.3)$$

No caso de variáveis reais, aos quais as probabilidades não são explícitas diretamente, opta pelo uso de estatísticas intervalares. Neste caso desejamos obter  $E(Y|x_1 < X < x_2)$ . Podemos obter, por exemplo, o valor da recuperação metalúrgica de carvão, dado que os valores de enxofre estejam entre 5% e 6%, por exemplo. O valor da esperança condicional considerando uma distribuição contínua das variáveis  $X$  e  $Y$  pode ser expressa pela equação (5.4)

$$E(Y|X) = \int_{-\infty}^{-\infty} y f_{Y|X}(y, x) dy \quad (5.4)$$

Onde  $f_{Y|X}(y, x)$  representa a função de densidade de probabilidade condicional. A figura 5.4 apresenta como calcular estatísticas condicionais considerando estatísticas intervalares. O histograma, ou a distribuição dos dados são consideradas dentro dos limites específicos  $x_1 < X < x_2$ , para um determinado tamanho da classe.

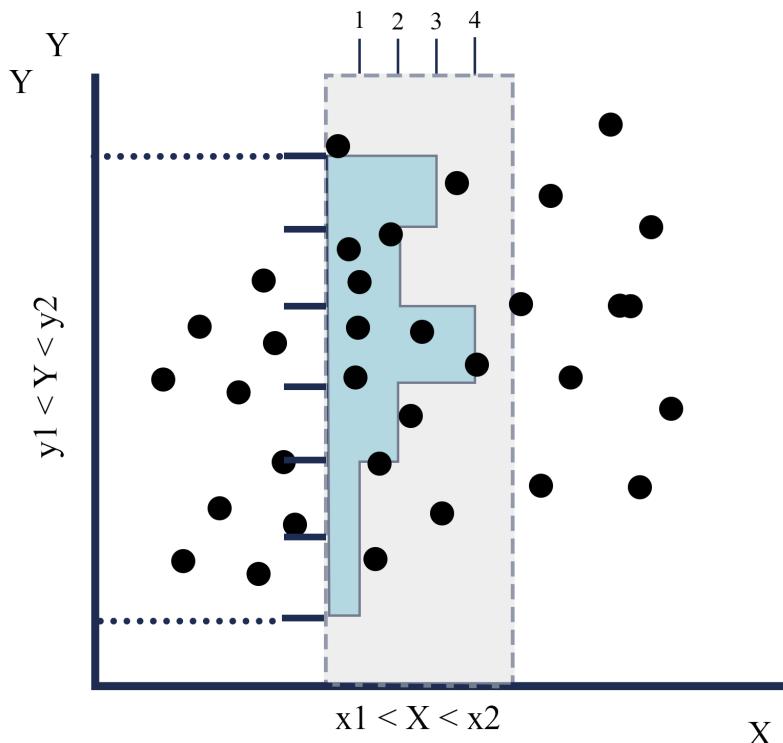


Figura 5.4: Demonstração do histograma condicional considerando um intervalo de  $x_1 < X < x_2$  e  $y_1 < Y < y_2$ . Podemos considerar

## 5.3 Ferramentas gráficas

### 5.3.1 Gráfico Q-Q plot

O gráfico q-q plot é uma ferramenta para uma primeira análise de diferentes distribuições de variáveis aleatórias. Para cada par conjugado são plotados os quantis de uma variável juntamente com outra. Variáveis que possuam distribuição semelhante tendem a apresentar um comportamento segundo uma reta  $y = x$ , de inclinação  $45^\circ$ .

Quando as variáveis apresentam a mesma forma, mas deslocamentos diferentes, ou seja, médias diferenciadas, o gráfico q-q plot apresenta o mesmo formato de uma reta, mas um deslocamento em sua abscissa. Quando as distribuições possuem formas semelhantes, mas variâncias diferentes, a distribuição tende a ter uma inclinação diferente. No entanto, quando distribuições possuem assimetrias e formas diferentes, o gráfico q-q plot tende a produzir uma convexidade diferente. A Figura 5.5 demonstra o gráfico q-q plot das variáveis Cobalto e Cádmio.

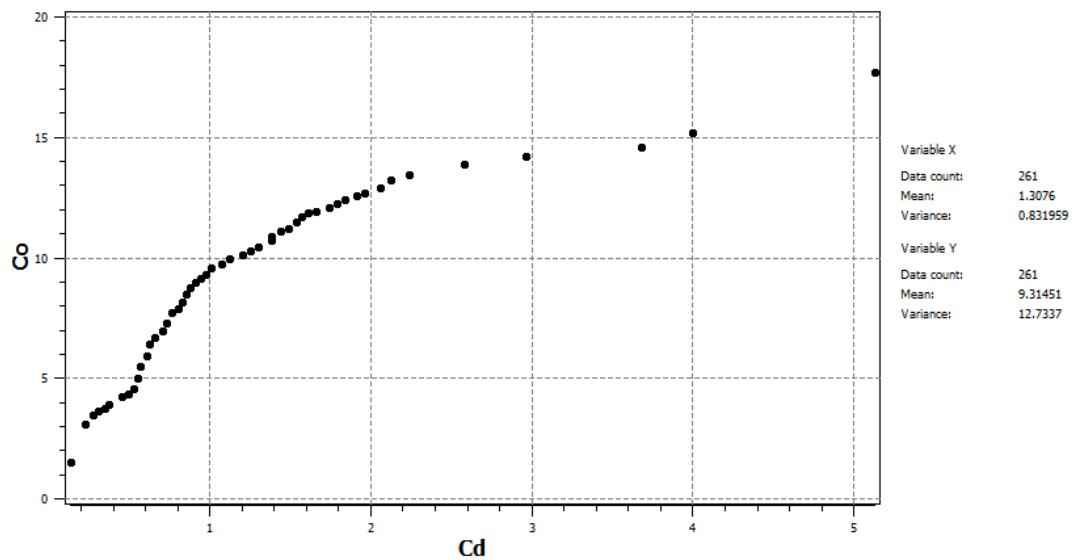


Figura 5.5: Gráfico QQ-Plot de Cobalto e Cádmio. Nota-se uma curvatura característica demonstrando pequena correspondência entre as duas populações. Cada ponto representa o mesmo quantil para cada variável

Nota-se na figura que o formato do q-q plot é convexo, demonstrando que as distribuições de dados seguem leis diferenciadas. A figura 5.6 demonstra a comparação entre os histogramas.

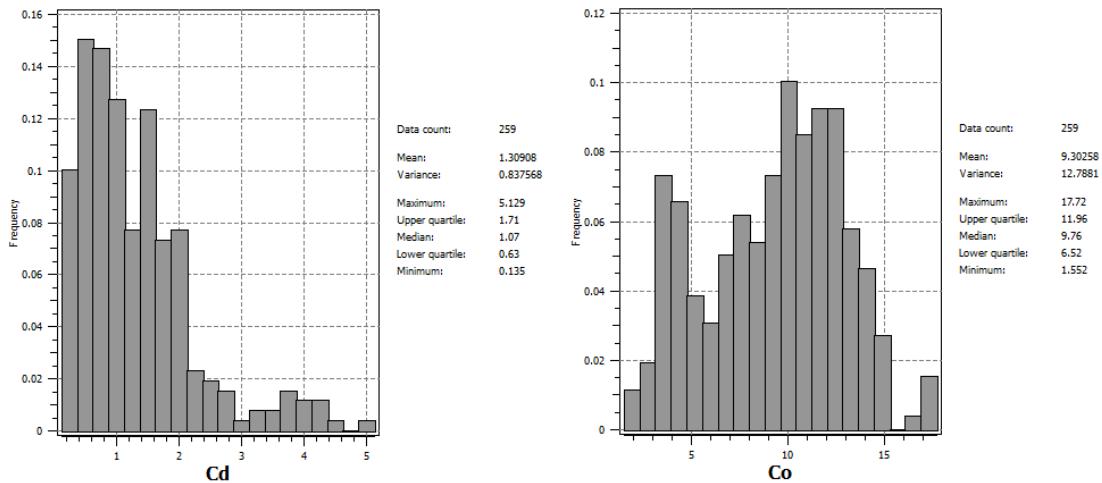


Figura 5.6: Diferenças entre os histogramas de Cádmio e Cobalto

**Proposição 5.3.1** *O gráfico q-q plot é uma alternativa para comparar distribuições de variáveis diferentes. A utilização da ferramenta, deve ser no entanto, utilizada com sabedoria. Valores outliers podem prejudicar a comparação entre as distribui-*

*ções, o que não significa que possam ser identificadas como possíveis distribuições semelhantes.*

Gráficos q-q plot podem ser utilizados não apenas entre amostras, mas também com uma combinação de uma variável amostrada e os quantis teóricos de uma distribuição. Uma das formas de se averiguar a normalidade de uma distribuição é comparar as amostras padronizadas  $Z_{pad} = (Z - \bar{x})/S$  com uma variável gaussiana padrão  $\phi(0, 1)$ . A figura 5.7 demonstra

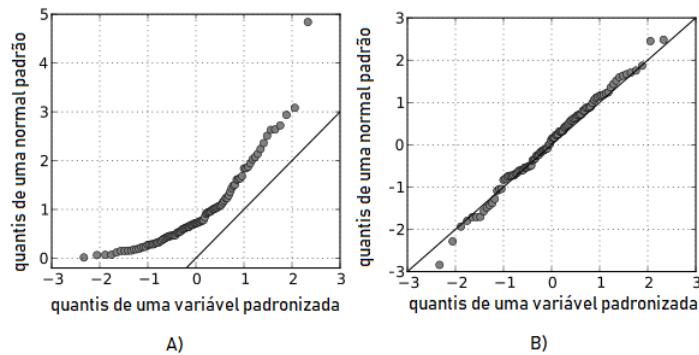


Figura 5.7: Gráfico da utilização do q-q plot para ajuste de uma distribuição. Quantis de uma amostra padronizada comparadas com quantil de uma distribuição gaussiana padrão. A) Mal ajuste. B) Bom ajuste

### 5.3.2 Gráfico p-p plot

Semelhante ao gráfico q-q plot temos o gráfico p-p plot. O gráfico de probabilidades implica nos pares conjugados que indicam a mesma probabilidade ( $Pr(Z < z), Pr(Y < z)) \forall z \in Z, Y$ ). A figura 5.8 demonstra o gráfico da variável Cobre pela de Cromo.

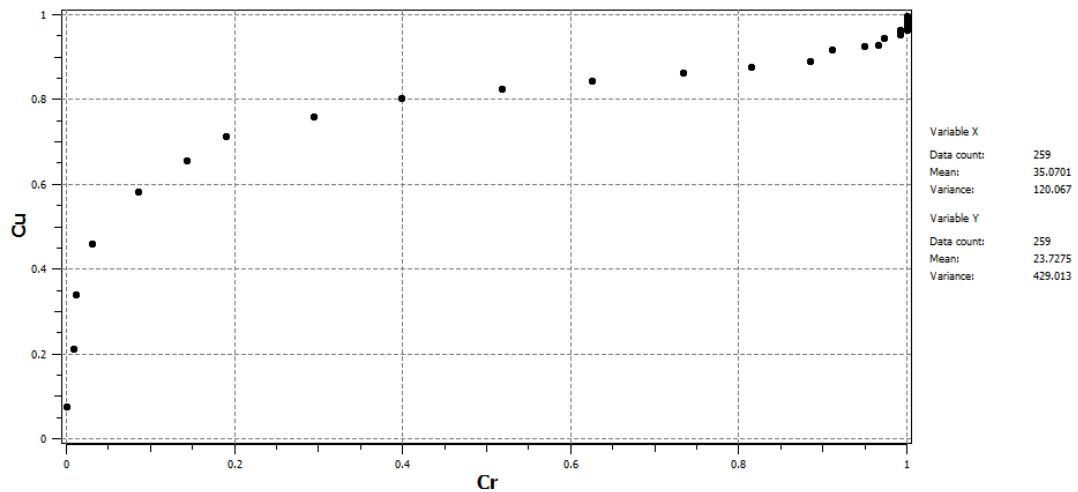


Figura 5.8: Gráfico PP-Plot de Cobre e Cromo. Nota-se uma curvatura característica demonstrando pouca correspondência entre as duas populações. Cada ponto representa o percentual acumulado para o mesmo valor da variável aleatória

Podemos notar que as diferenças demonstradas no gráfico p-p plot se reproduzem nas diferenças entre os histogramas de cobre e cromo na figura 5.9

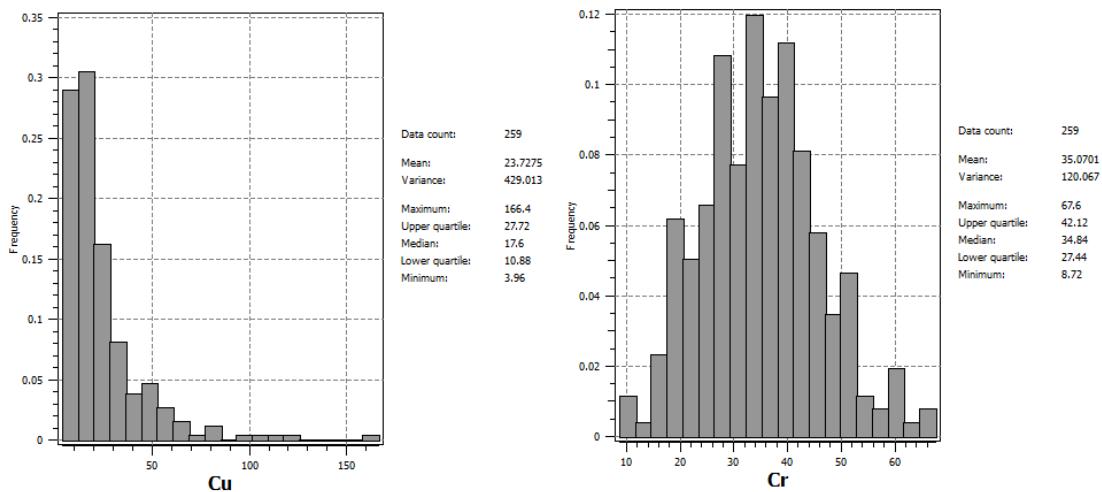


Figura 5.9: Diferenças entre os histogramas de Cobre e Cromo

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.

### 5.3.3 Gráfico de dispersão

O gráfico de dispersão apresenta dados de duas variáveis dispostos nos eixos cartesianos. Temos uma variável **preditora** (X) e uma variável **resposta** (Y). Os pares conjugados  $(x_i, y_i) \in X, Y$  representam pontos em um plano cartesiano.

**Proposição 5.3.2** *Uma das primeiras utilizações da regressão linear foi no estudo da importância de tendências entre gerações. Durante o período de 1893-1898, E. S. Pearson organizou uma coleção de  $n=1375$  alturas de mães do reino Unido abaixo de 65 anos e uma de suas filhas acima de 18. O interesse era computar o tamanho das mães ( $Mheight$ ) com o tamanho das filhas ( $Dheight$ ) como preditor. Se todas as mães possuirem tamanho igual suas filhas, os dados estarão dispostos em uma reta de inclinação  $45^{\circ}$ . A linearidade proposta pela dispersão identifica que mães mais altas geralmente possuem filhas mais altas. - Weisberg [2005]*

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 (5.10) demonstra um gráfico de dispersão entre a variável cromo e cobalto.

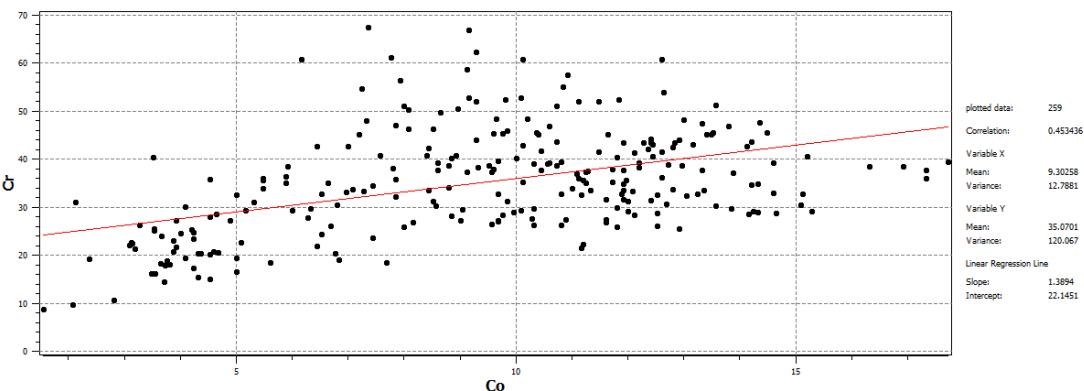


Figura 5.10: Gráfico de dispersão da variável Cromo e Cobalto. Nota-se dependência linear positiva entre as variáveis.

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. Neste caso possuímos uma **dependência linear negativa**

A Figura 5.11 demonstra os tipos de correlação lineares possíveis. Em 5.11 -a temos a correlação linear positiva em que o aumento da variável X aumenta o valor de Y, em 5.11 -b temos a correlação linear negativa em que o aumento do valor X tende a diminuir o valor de Y e em 5.11 -c temos um caso de independência entre as variáveis, tal que o aumento da variável X não altera o valor da variável Y.

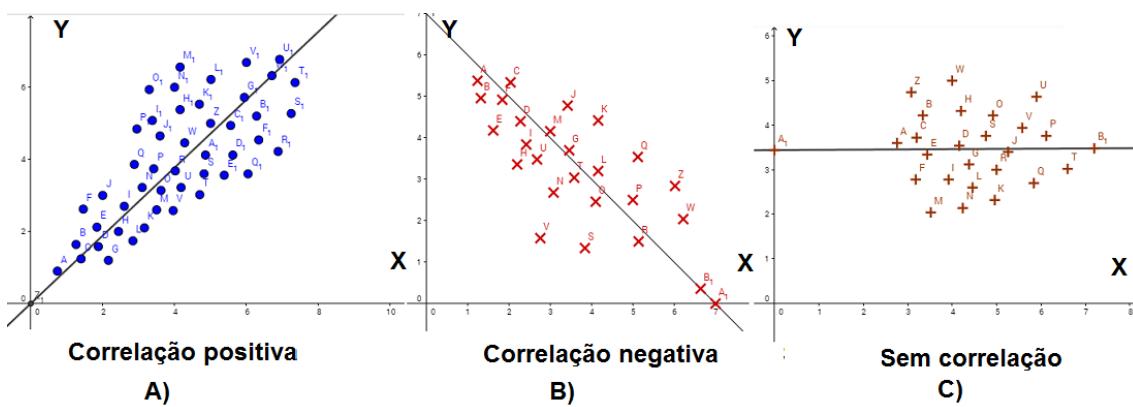


Figura 5.11: Figura demonstrando os tipos de correlação linear possíveis. A) Correlação linear positiva, B) Correlação linear negativa, C) Sem correlação

Os gráficos de dispersão também são uma boa medida para a visualização de valores outliers. A figura (5.12) 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.

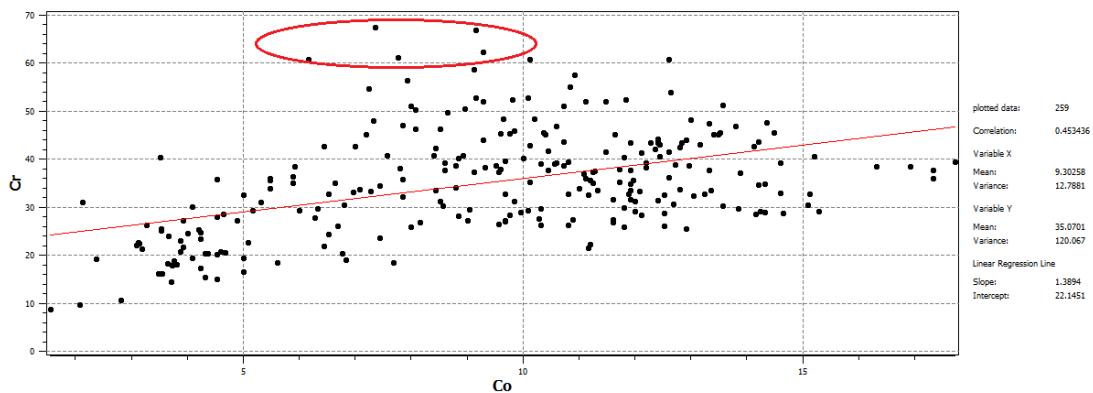


Figura 5.12: Gráfico de dispersão da variável Cromo e Cobalto demonstrando valores outliers. Círculo vermelho indica possíveis valores fora dos padrões das variáveis conjuntas

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.

## 5.4 Regressão linear

O modelo de regressão linear simples é aquele em que definimos uma dependência diretamente proporcional entre a variável resposta  $Y$  e preditora  $X$ . Podemos associar o valor esperado da variável resposta dado valores da variável preditora tal que  $E(Y|X = x) = \beta_0 + \beta_1 x$ . Note que  $E(Y|X = x)$  corresponde ao **valor médio da variável  $Y$  condicionado a um valor  $x$  da variável  $X$** , e  $\beta_0$  e  $\beta_1$ , também são o **intercepto da reta no eixo das abcissas e a tangente do ângulo de inclinação da reta**. Muitas pessoas acabam por não entender que a regressão linear pode não apresentar uma representação acurada da resposta dado um valor da variável preditora, porque o valor estimado pela regressão não é o valor da variável resposta, mas sim seu valor esperado condicionado. Muitas variáveis apresentam alta dispersão em torno de seus valores médios e podem não ser estimativas plausíveis. A solução da **regressão linear ordinária** geralmente advém do método dos mínimos quadrados. Considere  $\hat{y}_i = \hat{E}(Y|X = x)$  como um estimador para  $E(Y|X = x)$ , logo teremos que o resíduo pode ser demonstrado pela equação  $\hat{y}_i - y_i$  segundo a equação

*Demonstração.* Regressão linear pelo método dos mínimos quadrados

$$\begin{aligned}\epsilon_i &= \hat{y}_i - y_i \\ \epsilon_i^2 &= \hat{y}_i^2 + y_i^2 - 2\hat{y}_i y_i \\ \epsilon_i^2 &= (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2 + y_i^2 - 2(\hat{\beta}_0 + \hat{\beta}_1 x_i)y_i\end{aligned}$$

A soma dos erros quadráticos para cada  $i$

$$\sum_{i=0}^n \epsilon_i^2 = \sum_{i=0}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2 + \sum_{i=0}^n y_i^2 - \sum_{i=0}^n 2(\hat{\beta}_0 + \hat{\beta}_1 x_i)y_i$$

Tomando as derivadas parciais segundo os parâmetros:  $\hat{\beta}_0, \hat{\beta}_1$

$$\begin{aligned}1) \frac{\partial \sum_{i=0}^n \epsilon_i^2}{\partial \hat{\beta}_0} &= 2\hat{\beta}_0 n + 2\hat{\beta}_1 \sum_{i=0}^n x_i - \sum_{i=0}^n 2y_i = 0 \\ 2) \frac{\partial \sum_{i=0}^n \epsilon_i^2}{\partial \hat{\beta}_1} &= 2\hat{\beta}_1 \sum_{i=0}^n x_i^2 + 2 \sum_{i=0}^n \hat{\beta}_0 x_i - 2 \sum_{i=0}^n y_i x_i = 0\end{aligned}$$

■

A obtenção dos parâmetros pode ser facilmente encontrada isolando os termos das equações 1 e 2. Podemos então determinar

*Demonstração.* Obtenção dos parâmetros da regressão

$$\hat{\beta}_0 = \left( \sum_{i=0}^n y_i - \hat{\beta}_1 \sum_{i=0}^n x_i \right) / n$$

Substituindo em 2

$$\begin{aligned}\hat{\beta}_1 \sum_{i=0}^n x_i^2 + \sum_{i=0}^n \left( \sum_{j=0}^n y_j - \hat{\beta}_1 \sum_{j=0}^n x_j \right) x_i / n - \sum_{i=0}^n y_i x_i &= 0 \\ \hat{\beta}_1 \sum_{i=0}^n x_i^2 - \hat{\beta}_1 \bar{x} \sum_{i=0}^n x_i + \bar{x} \sum_{i=0}^n y_i - \sum_{i=0}^n y_i x_i &= 0 \\ \hat{\beta}_1 \left( \sum_{i=0}^n x_i^2 - \bar{x} \sum_{j=0}^n x_j \right) &= \sum_{i=0}^n y_i x_i - \sum_{j=0}^n y_j \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=0}^n y_i x_i - \sum_{j=0}^n y_j \bar{x}}{\sum_{i=0}^n x_i^2 - \bar{x} \sum_{j=0}^n x_j}\end{aligned}$$

■

O problema de regressão linear se torna então um problema de otimização ao encontrar o menor somatório dos desvios quadráticos. A figura (5.14) 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**. O resultado também é análogo ao **método de máxima verossimilhança** considerando a distribuição dos resíduos como gaussianos.

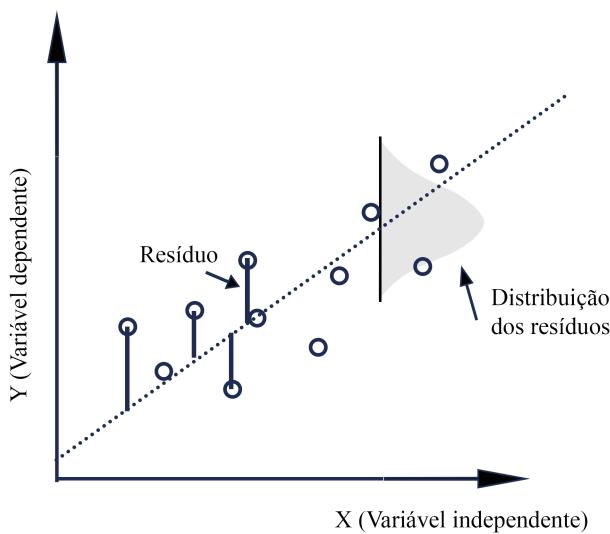


Figura 5.13: Explicação da regressão linear entre a variável independente X e a variável dependente Y. Barras verticais representando os desvios das amostras com o valor médio.

Como a regressão linear assume a minimização dos valores quadráticos, os parâmetros  $\beta_0$  e  $\beta_1$  podem ser fortemente afetados por valores outliers. As propostas mais modernas de regressão prevem que a regressão seja utilizada apenas em parte dos dados, e avaliada com outra quantidade dos dados. Os dados utilizados para a estimativa dos parâmetros é chamada de Assim conseguimos avaliar a qualidade da predição de acordo com a dispersão destes resíduos.

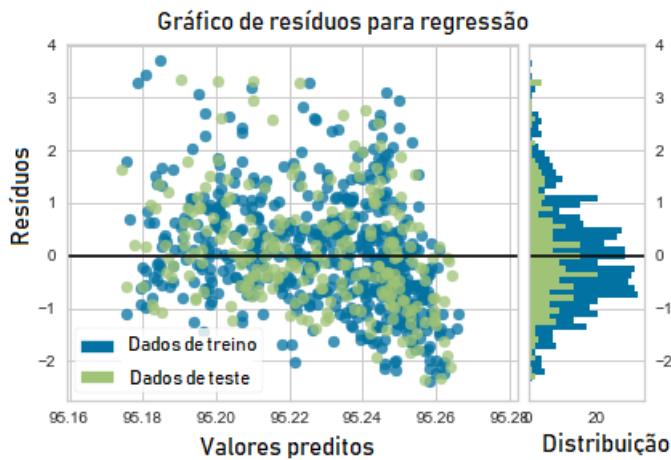


Figura 5.14: Gráfico da dispersão de resíduos nos bancos de dados de treino e teste. Dados de treino utilizados para determinar os parâmetros da regressão e dados de teste para avaliar as diferenças do modelo de predição

Na geoestatística utilizamos estimadores lineares, semelhantes ao processo de regressão linear. É observado que para aplicações na mineração que utilizem os métodos geoestatísticos clássicos, o erro do valor estimado é ligeiramente diferente de uma distribuição gaussiana, para problemas lineares e estacionários. Temos uma maior confiabilidade do valor esperado estimado utilizando krigagem ordinária do que utilizando regressão linear ordinária. Na verdade, o método de regressão linear ordinária é pouco usual nos dias atuais, considerando os diferentes modelos possíveis e robustos (menos influenciados pelos valores outliers), no entanto, ainda é um método muito popular pela sua simplicidade e facilidade de aplicação.

**R** *Em aplicações da mineração, a distribuição dos erros da função aleatória são geralmente simétricos com um crescimento mais pronunciado na moda e caudas alongadas do que para distribuições normais com mesma média e variância. Então em relação a uma distribuição normal, há menos erros na região próxima ao valor estimado e mais erros nas caudas. - Journel and Huijbregts [1978]*

## 5.5 Intervalo de segurança para a regressão linear

A determinação do modelo de regressão consiste em estimar parâmetros  $\beta_0$  e  $\beta_1$  para econtrarmos o valor estimado  $\hat{y}_i = E(Y|X = x_i)$ . No entanto, se a nuvem de pontos determinada pela regressão for esparsa, o valor  $\hat{y}_i$  não possui capacidade preditiva e pode encontrar-se dentro de limites amplos. Considerando que a distribuição dos resíduos seja normalmente distribuída, podemos encontrar os limites da regressão

para bandas superiores e inferiores, determinando assim a confiabilidade desta reta regredida. A equação (5.5) demonstra como podem ser calculadas as bandas de incerteza da regressão de acordo com um nível de significância estipulado.

$$\hat{y}_i \pm t_{(n-2,p)}^* s_y \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}} \quad (5.5)$$

Em que  $x_i$  é o valor da variável X,  $t^*$  é o valor da distribuição de t-student para um grau de liberdade igual a  $n - 2$  e nível de significância p, enquanto  $s_y$  pode ser demonstrado segundo a equação (5.6)

$$s_y = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} \quad (5.6)$$

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 (5.15) demonstra o intervalo de segurança para o valor regredido.

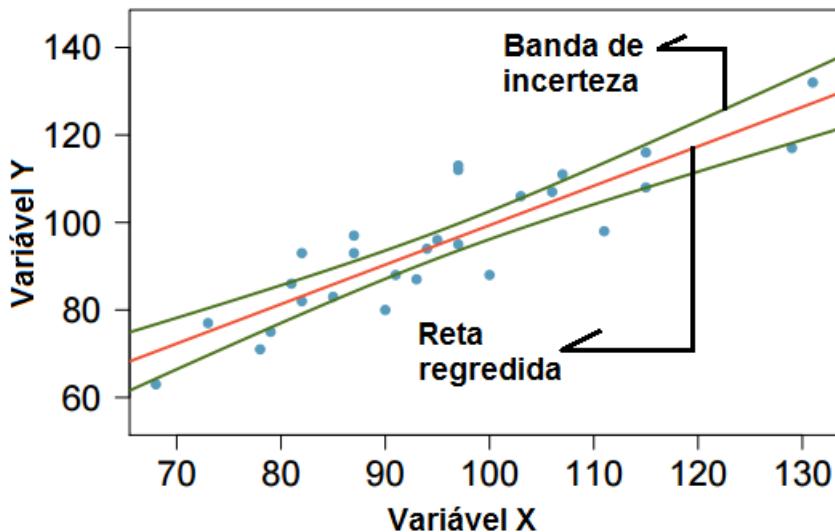


Figura 5.15: Demonstração do intervalo de confiança para a regressão linear. Banda de incerteza adicionada como limite inferior e superior dado pela equação (5.5)

Nota-se que as bandas apesar de acompanharem o valor de regressão linear não são retas, apresentando um maior estreitamento na região mediana da dispersão. A confiabilidade do centro de dispersão da reta regredida é sempre maior.

**Proposição 5.5.1** Os intervalos de confiança para a regressão linear estipulam que os resíduos seguem uma distribuição gaussiana. Isto porém, pode não se apresentar na prática. A única garantia que temos para que este resíduo seja considerado gaussiano, é se por ventura, a distribuição dos dados segue uma lei de probabilidades **multigaussiana**. Este é o pressuposto de técnicas mais avançadas de geoestatística não-linear. As bandas de incerteza devem ser consideradas como uma alternativa para verificar dados discrepantes, mas não como uma métrica de decisão na detecção de valores outliers.

## 5.6 Regressão linear múltipla

Quando pensamos em apenas uma variável preditora, a determinação de  $\hat{y}_i$  se limita a encontrar o valor de  $E(Y|X = x_i)$ . Porém quando múltiplas variáveis são relacionadas o problema se torna encontrar  $E(Y|X^1 = x_i^1, X^2 = x_i^2, \dots, X^n = x_i^n)$ . 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 preditoras e uma variável resposta

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \dots + \hat{\beta}_n x_i^n = \sum_{j=0}^p \hat{\beta}_j x_i^j \quad (5.7)$$

Onde  $x_i^0$  é sempre igual a 1 para  $j$  variáveis de 0 a  $p$ . O problema se resume a encontrar  $\hat{\beta}_p$  parâmetros que aproximem melhor a combinação dos valores das variáveis  $x_i^p$  de  $\hat{y}_i$ . Podemos definir o problema de regressão linear múltipla a partir de sua forma matricial pela equação (5.8)

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ 1 & x_2^1 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \quad (5.8)$$

Ou de forma simplificada pela equação (5.9)

$$\bar{Y} = \bar{X}\bar{\beta} \quad (5.9)$$

Onde  $\bar{Y}$  representa o vetor da variável resposta,  $\bar{X}$  a matriz das variáveis preditoras e  $\bar{\beta}$  os parâmetros. A obtenção dos parâmetros a partir da regressão múltipla

pode ser facilmente encontrado através de operações matriciais.

*Demonstração.* Obtenção dos parâmetros da regressão

$$\begin{aligned}\bar{Y} &= \bar{X}\bar{\beta} \\ \bar{X}\bar{Y} &= \bar{X}\bar{X}\bar{\beta} \\ (\bar{X}\bar{X})^{-1}\bar{X}\bar{Y} &= \bar{\beta} \\ (\bar{X}\bar{X})^{-1}\bar{X}\bar{Y} &= \bar{\beta} \\ X^\dagger\bar{Y} &= \bar{\beta} \\ \text{tal que } X^\dagger &= (\bar{X}\bar{X})^{-1}\bar{X}\end{aligned}$$

■

Em que  $X^\dagger$  também é chamada de pseudo-inversa de  $X$ . Se no caso da regressão ordinária simples obtínhamos um valor regredido a partir da minimização do resíduo de uma variável, neste momento obtemos o resíduo a partir de uma combinação de múltiplas variáveis. O erro quadrático pode ser obtido a partir da equação (5.10).

$$\sum_i \epsilon_i^2 = \sum_i \left( y_i - \sum_{j=0}^p \hat{\beta}_j x_i^j \right)^2 \quad (5.10)$$

A obtenção dos parâmetros  $\beta_p$  pode ser calculado a partir das técnicas de minimização dos resíduos, formando um sistema de  $p$  derivadas parciais. Um dos grandes problemas da regressão linear múltipla é o fato de que as grandezas de variáveis diversas podem ser diferentes e impactar de forma diferenciada nos pesos da regressão. Esse problema de dimensão geralmente pode ser minimizado se padronizarmos as variáveis como visto no capítulo 4. Outra questão envolvendo a regressão múltipla é o fato de que valores outliers conseguem ser ainda mais prejudiciais que a regressão linear ordinária, afetando muito a estimativa dos pesos. Uma tentativa de excluir certos efeitos de valores discrepantes é introduzir uma constante adicional chamada de **regularização** ( $\lambda$ ) multiplicando os valores dos parâmetros  $\beta_p$ .

## 5.7 Coeficiente de correlação

Observamos na seção anterior que se duas variáveis são dependentes, podemos assumir que existe uma probabilidade  $P(Y|X = x)$ . No entanto, as probabilidades condicionais parecem não fornecer um quadro geral da dependência de uma variável

aleatória  $Y$ , pois precisamos saber qual valor a variável  $X$  deve assumir. É necessário ter uma métrica para avaliarmos o quanto forte ou fraca é a dependência entre as variáveis. Imagine o caso onde temos um depósito hidrotermal de ouro, associado principalmente a rochas magmáticas sulfetadas. Se existir uma alta dependência do conteúdo de ouro com o de enxofre, podemos assumir que o conhecimento de uma variável auxiliará no conhecimento da outra. Porém valores pequenos de teor de enxofre podem ser menos dependentes do teor de ouro do que para altos valores do teor de enxofre. Esta discrepância dado alguns limites pode ser favorável para o uso de probabilidades condicionais nas caudas da distribuição de enxofre, mas não garante uma visão geral da dependência linear entre estas variáveis.

**R** *Tanto na natureza como em vários problemas de engenharia nos deparamos com a dependência entre diferentes variáveis. Em muitos casos estas dependências podem ser modeladas linearmente. Em outros casos, quando conhecemos propriedades físicas relacionáveis, podemos utilizar transformações lineares, capazes de transformar modelos não lineares em lineares.*

No capítulo 3 observamos a correlação como uma medida de dependência entre variáveis aleatórias. A covariância teórica pode ser estimada a partir de sua covariância experimental pela equação (5.11)

$$\hat{Cov}_{X,Y} = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (5.11)$$

Onde  $\bar{x}$  e  $\bar{y}$  são as médias aritméticas entre as variáveis X e Y para um número de amostras n. A covariância experimental pode ser muito bem comparada ao produto escalar obtido pela multiplicação de dois vetores  $X * Y = \|X\| \|Y\| \cos(\theta)$ , em que  $\cos(\theta)$  é chamado de cosseno diretor da projeção de  $X$  em  $Y$ . Quando a projeção do vetor  $X$  corresponde ao vetor  $Y$  temos o máximo valor possível, no entanto, quando estes vetores são perpendiculares temos um valor de  $\cos(90^\circ) = 0$ , e portanto  $\hat{Cov}_{X,Y} = 0$

A covariância é uma medida muito susceptível a valores outliers, pois como X e Y podem ter unidades discrepancytes, o produto das duas pode ser mais influenciado por aquela variável que apresentar maiores valores. Imagine que procuremos a relação entre a massa em quilos dos testemunhos e o teor de um elemento químico. Como a massa dos testemunhos poderá variar de valores acima da unidade, como por exemplo 12kg e os teores apenas com valores decimais, como 0.12, a importância dada para as variações do peso serão maiores. Isto torna a covariância pouco comparativa, apenas se considerarmos a normalização das variáveis. A alternativa para

isto é normalizarmos a covariância pelos desvios padrões das respectivas variáveis, o que gera o **coeficiente de correlação de Pearson** (5.12).

$$\rho_{X,Y}^p = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (5.12)$$

Os valores de  $\rho_{X,Y}$  podem variar de -1 a 1, sendo 1 quando apresenta correlação positiva perfeita e -1 quando apresenta correlação negativa perfeita. Quando  $\rho_{X,Y}^p = 0$  temos um indicativo de independência entre as variáveis tal que  $\hat{Cov}_{X,Y}$  é igual a 0. Podemos obter a relação entre o coeficiente de Pearson e a correlação pela equação (5.13)

$$\rho_{X,Y}^p = \frac{\hat{Cov}_{X,Y}}{\sqrt{S_x^2 S_y^2}} \quad (5.13)$$

Onde  $\hat{Cov}_{X,Y}$  representa a covariância estimada, e  $S_x^2$  e  $S_y^2$  as variâncias experimentais das amostras x e y. Para que consigamos calcular o valor do coeficiente de variação e da correlação, precisamos que existam valores correspondentes tanto para as amostras x e y, ou seja, precisamos de dados **colocados**. Se por ventura houverem dados faltantes, não conseguiremos calcular a covariância ou o coeficiente de correlação.

Em alguns casos não desejamos observar a dependência entre os valores da variável, mas precisamos saber qual é a dependência da **ordem** dos dados. Para isto realizamos uma medida chamada de **rank ou posto**, que representa a ordem de uma amostra em seu conjunto.

**Definição 5.7.1 — Rank ou posto.** *Um rank ou posto é a ordem crescente de uma amostra em seu respetivo conjunto. Por exemplo, um conjunto de amostras com valores  $x = 3, 41, 2, 57, 8, 9, 6$ , possuirá um rank  $R_x$  tal qual  $R_x = 2, 6, 1, 7, 4, 5, 3$*

O chamado **coeficiente de correlação de Spearman** nada mais é que a correlação de Pearson considerando seus respectivos postos.

$$\rho_{X,Y}^s = \frac{\hat{Cov}_{R_x, R_y}}{\sqrt{S_{R_x}^2 S_{R_y}^2}} \quad (5.14)$$

Enquanto a correlação de Pearson é uma medida linear de dependência, o coeficiente de Spearman é uma medida não linear, demonstrando a correlação entre

crescimentos e descrescimentos das variáveis independente dos valores dos dados. Uma função parabólica tal que  $Y = \beta_1 X^2 + \beta_0$  apresentará um valor de coeficiente de correlação de Pearson muito baixo, porém um valor de Coeficiente de Spearman igual a 1.

## 5.8 Exercícios

**Exercícios 5.1** Os dados da tabela abaixo representam valores de Au e cobre medidos concumitamente nos mesmos testemunhos de sondagem. Com estes dados, pede-se:

- Determine a covariância dos dados
- Determine o coeficiente de correlação.
- As amostras são dependentes positivamente ou negativamente?
- 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

**Exercícios 5.2** Os dados da tabela abaixo representam valores de X e Y. Faça um gráfico de dispersão e determine o par de valor outlier 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



## 6. Métodos clássicos e desagrupamento

*Difficultés rencontrées dans le développement d'une Géostatistique linéaire. A cette émergence lente et difficile, nous apercevons plusieurs sortes de raison, les unes historiques, d'autres simplement psychologiques, et d'autres encore qui correspondent a des problèmes de fond, à de véritables difficultés méthodologiques, ou épistémologiques qui n'ont été vraiment élucidées qu'à la fin des années 60 ou au début des années 70*

*G. Matheron*

### 6.1 Introdução

Apesar de antiga, a geoestatística se iniciou como um alternativa para tentativas de avaliação de depósitos minerais antes da década de 70. Os métodos chamados de clássicos eram relativamente eficientes para condições de depósitos minerais mais homogêneos e de classificação estatística dentro dos grupos considerados regulares. No entanto, devido a intensa atividade industrial humana, é necessário aproveitarmos cada vez mais depósitos minerais complexos. Diferentemente dos métodos

geoestatísticos, os métodos convencionais baseiam-se apenas na distribuição geométrica entre as amostras e não na correlação e dependência entre variáveis aleatórias. Apesar de ultrapassados, os métodos de avaliação clássica ainda são utilizados em depósitos de baixa variabilidade e com quantidades de amostras muito baixas, como por exemplo, depósitos estratiformes de argila ou areia. Em alguns casos bem específicos os métodos clássicos podem até mesmo apresentar resultados superiores aos métodos geoestatísticos seguindo o princípio da navalha de Occan.

**Definição 6.1.1 — Navalha de Occan.** *"A navalha de Occam é um princípio lógico atribuído ao filósofo medieval William de Occam. O princípio estabelece que não se deve assumir mais suposições do que necessário. Também é chamado de princípio da parsimônia. Este princípio envolve todo a modelagem científica e construção de teorias. Ele nos incentiva a escolher de um grupo de modelos equivalentes para um dado fenômeno aquele mais simples. Para todo modelo, a navalha de Occam nos ajuda a 'cortar' aqueles conceitos, variáveis ou construções que não conseguem realmente explicar o fenômeno. Ao fazer isso, o desenvolvimento do modelo se torna bem mais simples, e há menos chances de introduzir inconsistências, ambiguidades ou redundâncias"* -[Heylighen \[1997\]](#)

Veja bem que adotar os métodos clássicos em detrimento da geoestatística utilizando a navalha de Occan só possui bons resultados em dois casos. No primeiro o problema é tão simples e o depósito mineral tão homogêneo e pouco amostrado, que se torna factível o uso de um modelo extremamente simplificado. No segundo caso temos um problema tão complexo e variável que se torna impossível encontrar aparentemente um padrão qualquer no fenômeno, sendo mais fácil adotar valores médios pela distância do que realmente utilizar um método geoestatístico refinado. Segundo o professor [Yamamoto \[2001\]](#), os métodos chamados de clássicos baseiam-se no princípio da **interpretação**, aos quais determinam valores a partir de duas amostras contíguas. É possível a partir da disposição das amostras encontrar valores estimados. Estes princípios segundo o professor são:

1. Mudança gradual ou lei de função linear
2. Pontos mais próximos ou esfera de igual influência
3. Generalização ou empírico

### **6.1.1 Princípio da mudança gradual**

O princípio da mudança gradual indica que uma mudança de uma propriedade acontece de forma contínua de uma amostra pontual  $P_1$  até uma amostra pontual  $P_2$ .

Pelo princípio da navalha de Occan, a função utilizada para realizar esta transição geralmente é uma variação linear. Dada uma propriedade como o teor  $T$  a ser estimada a partir das propriedades  $T_1$  e  $T_2$  de dois pontos amostrais no espaço, com respectivas distâncias  $D_1$  e  $D_2$  da origem, realizamos a interpolação deste valor a partir de sua distância  $D$  entre os pontos pela equação (6.1).

$$T = T_1 + \frac{(D - D_1)(T_2 - T_1)}{(D_2 - D_1)} \quad (6.1)$$

A figura 6.1 apresenta o resultado geométrico da interpolação linear realizada entre os pontos amostrais.

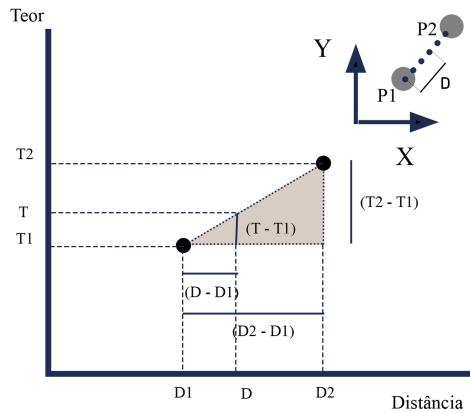


Figura 6.1: Princípio da mudança gradual de um teor para duas amostras no espaço  $P_1$  e  $P_2$ , variação linear dos teores para uma distância  $D$  considerada.

### 6.1.2 Princípio dos pontos mais próximos

O princípio dos pontos mais próximos assume que o valor interpolado é igual ao valor da amostra mais próxima dele. Este processo também é chamado de **vizinho mais próximo**. Dado um conjunto de amostras no espaço  $P = \{P_1, P_2, P_3, \dots, P_n\}$  com propriedades  $T = \{T_1, T_2, T_3, \dots, T_n\}$ , com respectivas posições espaciais segundo um eixo cartesiando  $(x, y, z)$ , o valor de uma propriedade  $T_0$  para uma amostra  $P_0$  no espaço assume o valor

$$T_0 = T_i | i \rightarrow \min(\|P_0, P_i\|), \forall i \in [1, n] \quad (6.2)$$

Em que  $\|P_0, P_i\|$  representa a distância euclidiana entre o par conjugado de pontos no espaço para  $n$  amostras,  $\min()$  representa o mínimo valor. A figura 6.2

apresenta o princípio dos pontos mais próximos. Cinco pontos amostrais são apresentados  $P_1$ ,  $P_2$ ,  $P_3$ ,  $P_4$  e  $P_5$ . Como a distância mínima entre o ponto a ser amostrado  $P_0$  até o ponto amostral mais próximo é  $D_2$ ,  $T_0$  recebe o valor de  $T_2$ .

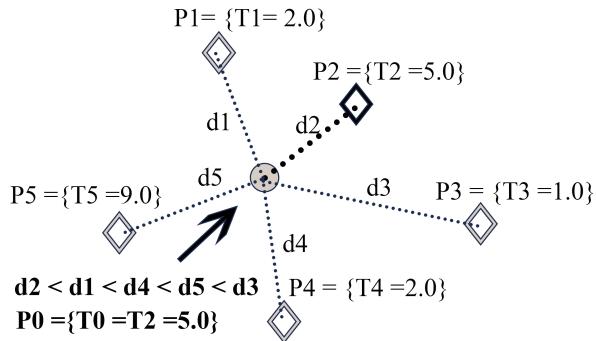


Figura 6.2: Princípio dos pontos mais próximos. Conjunto de pontos amostrais  $P_1$ ,  $P_2$ ,  $P_3$ ,  $P_4$ ,  $P_5$  para um ponto amostral estimado  $P_0$ . Como a menor distância euclidiana entre o ponto  $P_0$  é o ponto  $P_2$ , assumimos que a propriedade  $T_0 = T_2$ .

### 6.1.3 Princípio da generalização

Segundo critérios geológicos é possível realizar a extração de uma dada propriedade segundo a continuidade do depósito mineral. Este princípio é justificado nas fases inciais de pesquisa para ajustar uma dada tendência das propriedades do depósito mineral. A figura 6.6 demonstra a extração do teor a partir do conhecimento de uma falha geológica entre os pontos amostrais  $P_1$  e  $P_2$ . A partir da atitude da camada é estipulado uma variação segundo os teores extrapolados.

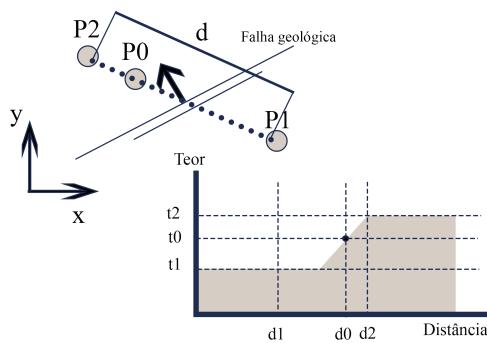


Figura 6.3: Princípio da generalização e extensão dos teores a partir de um critério geológico. Falha geológica observada entre os pontos amostrais  $P_1$  e  $P_2$ . Ponto estimado  $P_0$  é determinado a partir de diferenças entre os teores e inclinação da camada.

## 6.2 Composição

Para realizar a análise geoestatística é necessário prover de amostras com mesmo **suporte**. Como testemunhos de sondagem podem representar fragmentos de tamanho distintos, é necessário realizar a **regularização** dos tamanhos das amostras. Isto significa que ao invés de tomarmos as propriedades referentes a apenas os fragmentos dos testemunhos de sondagem (litotipo, teores, propriedades físicas, etc.), tomamos a amostra como um valor de média ponderada entre os diversos tamanhos de fragmentos dos testemunhos. Observe a figura 6.4. Os fragmentos do testemunho de sondagem são respectivamente os valores  $\{l_1, l_2, l_3, l_4\}$ . Para considerarmos este testemunho de sondagem como medidas representativas para a geoestatística consideramos duas composições de tamanho  $\{C_1, C_2\}$ , em que  $\{x_1, x_2\}$  representam as respectivas partes dos fragmentos  $\{l_1, l_2\}$  em  $C_1$ . Se  $t_1, t_2, t_3, t_4$  são propriedades aditivas como os teores dos fragmentos, podemos encontrar o valor da propriedade da composição  $t_{C1}$  pela relação (6.3)

$$t_{C1} = \frac{(x_1/l_1)t_{l_1} + (x_2/l_2)t_{l_2}}{(x_1/l_1) + (x_2/l_2)} \quad (6.3)$$

Em que  $(x/l)$  representa a proporção de um dado fragemento dentro da composta. Podemos generalizar o caso da composição para n fragmentos de acordo com a relação (6.4)

$$t_C = \frac{\sum_{i=0}^n (x_i/l_i)t_{l_i}}{\sum_{i=0}^n (x_i/l_i)} \quad (6.4)$$

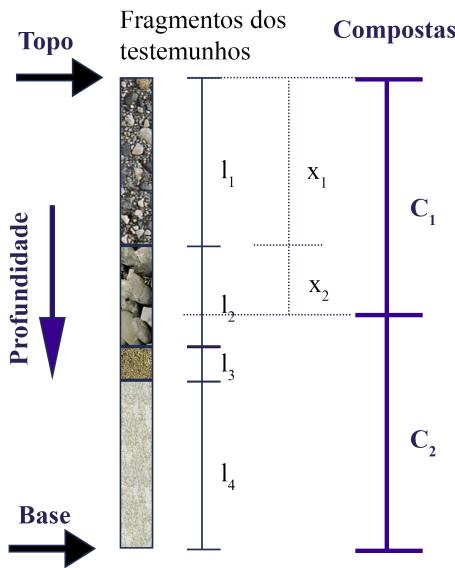


Figura 6.4: Composição realizada em testemunho com 4 fragmentos  $\{l_1, l_2, l_3, l_4\}$  e tamanhos de composta igual  $\{C_1, C_2\}$ .  $\{x_1, x_2\}$  representam respectivamente as proporções dos fragmentos  $\{l_1, l_2\}$  dentro da composta  $C_1$ .

A escolha do tamanho da composição do testemunho geralmente está associada ao comprimento da manobra. Em alguns casos quando são realizados furos com manobras de tamanho diferenciado pode se optar pela moda dos valores de manobra.

**Proposição 6.2.1** *Devemos lembrar antes de realizar a composição de uma dada propriedade se ela pode ser caracterizada como uma variável aditiva, tal como teores e massas. A composição de testemunhos de sondagem considerando propriedades não aditivas deve ser realizada a partir dos estimadores adequados para as tendências centrais. Por exemplo, se considerarmos a velocidade de propagação de um pulso sísmico nos fragmentos, sabemos que a média correta de velocidade é a **harmônica** e não a média **aritmética**.*

### 6.3 Composição em seções verticais

É comum durante o processo de estimativa, o geólogo realizar seções interpretadas de um depósito mineral, considerando os aspectos estruturais do depósito, o controle geológico entre outras características que formam um **modelo geológico**. Diferentemente de um **modelo estatístico** que prevê o conhecimento de variáveis aleatórias do depósito mineral, o modelo geológico é uma representação dos diferentes litotipos dispostos no espaço, e nem sempre podem ser correlacionados com suas devidas propriedades. Para incorporar os valores das propriedades nas interpretações geológicas

podemos fazer um "preenchimento" das seções a partir de valores médios obtidos nelas. Este processo segue o princípio da generalização tomando o valor médio de uma seção interpretada como valor médio dos testemunhos contidos dentro daquela seção, ou daqueles bem próximos a seção interpretada. Observe a figura 6.5, temos 6 compostas realizadas em furos  $\{F_1, F_2, F_3, F_4, F_5, F_6\}$  correspondendo a área cinza da seção interpretada. Os valores de teores são respectivamente  $\{t_1, t_2, t_3, t_4, t_5, t_6\}$  de comprimentos  $\{l_1, l_2, l_3, l_4, l_5, l_6\}$

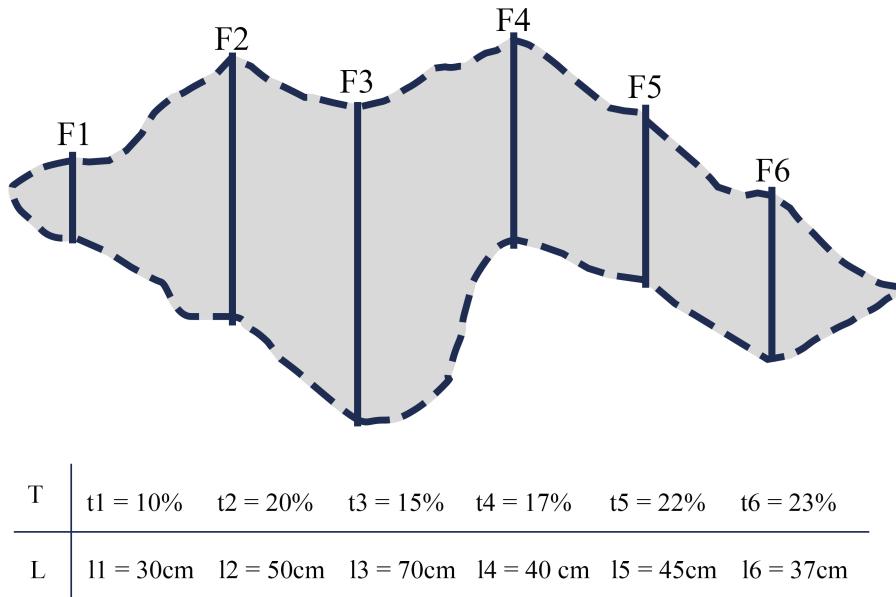


Figura 6.5: Composição realizada em testemunho com 4 fragmentos  $\{l_1, l_2, l_3, l_4\}$  e tamanhos de composta igual  $\{C_1, C_2\}$ .  $\{x_1, x_2\}$  representam respectivamente as proporções dos fragmentos  $\{l_1, l_2\}$  dentro da composta  $C_1$ .

Para obtermos o valor médio da seção podemos realizar a seguinte operação de composição (6.5)

$$t_C = \frac{\sum_{i=1}^6 t_i l_i}{\sum_{i=1}^6 l_i} = \frac{10 * 30 + 20 * 50 + 15 * 70 + 17 * 40 + 22 * 45 + 23 * 37}{30 + 50 + 70 + 40 + 45 + 37} = 17.91\% \quad (6.5)$$

## 6.4 Determinação de volumes

Um dos valores mais importantes obtidos na produção mineral é o volume do material a ser extraído, esse processo também é chamado de **cubagem**. A mineração consiste em movimentar um grande volume de rochas de diferentes litotipos. Infelizmente as informações obtidas das rochas são descritas apenas por amostras de pequeno volume e extensão, como em testemunhos de sondagem, durante as primeiras etapas da pesquisa mineral. Em alguns casos é possível obter trincheiras ou verificar as estruturas geológicas em painéis de mina subterrânea. Estes volumes estimados são geralmente obtidos pela interpretação geológica de um profissional engenheiro de minas ou geólogo, capaz de entender os processos de gênese destes depósitos minerais. Uma das alternativas consiste em realizar seções verticais contendo informações dos furos de sondagem e realizar a extrapolação de volumes a partir da extensão das áreas interpretadas. Dado uma série de seções  $S$  de área  $A$ , o volume entre seções pode ser obtido a partir da equação (6.6)

$$V = \sum_{i=1}^{n-1} D_i(A_i + A_{i+1})/2 \quad (6.6)$$

Onde  $D_i$  é a distância entre as seções  $A_i$  e  $A_{i+1}$  para  $n$  seções consideradas. Os volumes obtidos pelas seções das pontas é obtido a partir de extrapolação dado uma distância considerada aceitável pelo geólogo, baseando-se em critérios de continuidade. A figura 6.6 apresenta a interpolação realizada a partir da interpretação geológica das seções e extrapolação destas para obtenção do volume.

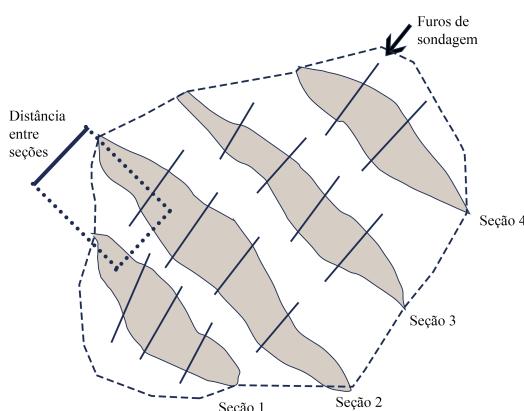


Figura 6.6: Obtenção dos volumes dos corpos geológicos a partir de extrapolação de seções verticais interpretadas. Cinco seções consideradas e os respectivos furos de sondagens utilizados para interpretar os volumes.

**Proposição 6.4.1** Apesar de ser um método antigo de avaliação de volumes, as interpretações de seções ainda são utilizadas na prática. Apesar da dificuldade operacional deste método com a digitalização de seção por seção, permite a geólogos e engenheiros de minas incorporarem informações importantes para o modelo, como adição de estruturas geológicas, possíveis zonas de alteração, entre outras características típicas do depósito analisado. Métodos recentes como a **modelagem implícita** auxiliam muito na velocidade de produção dos modelos, mas geralmente precisam de um cuidado maior ao incorporar outras informações após serem gerados

## 6.5 Inverso do quadrado da distância - IQD

Um dos interpoladores mais antigos conhecidos é o uso do inverso do quadrado da distância. A justificativa da utilização deste ponderador é que muitos problemas físicos ocorrem a partir de leis de decaimento quadráticas. O uso de outras potências também pode ser utilizado, mas a medida que esta potência cresce, os valores mais próximos das regiões estimadas tendem a ser mais valorizados, reduzindo a suavização da interpolação. Dada uma propriedade  $t$  como o teor de uma amostra  $P$ , situada a uma distância  $d$  do centroide de uma célula estimada, a média estimada desta célula pode ser calculada pela equação (6.7):

$$t_m = \frac{\sum_{i=1}^n \frac{1}{d_i^p} t_i}{\sum_{i=1}^n \frac{1}{d_i^p}} \quad (6.7)$$

Em que  $p$  é o grau do polinômio considerado para  $n$  amostras situadas nas redondezas da célula estimada. Se o número de amostras é grande, o cálculo dos ponderadores geralmente é realizado apenas com os valores mais próximos, para isto utilizando um algoritmo que filtre segundo a proximidade das amostras da célula estimada. Observe a figura 6.7.

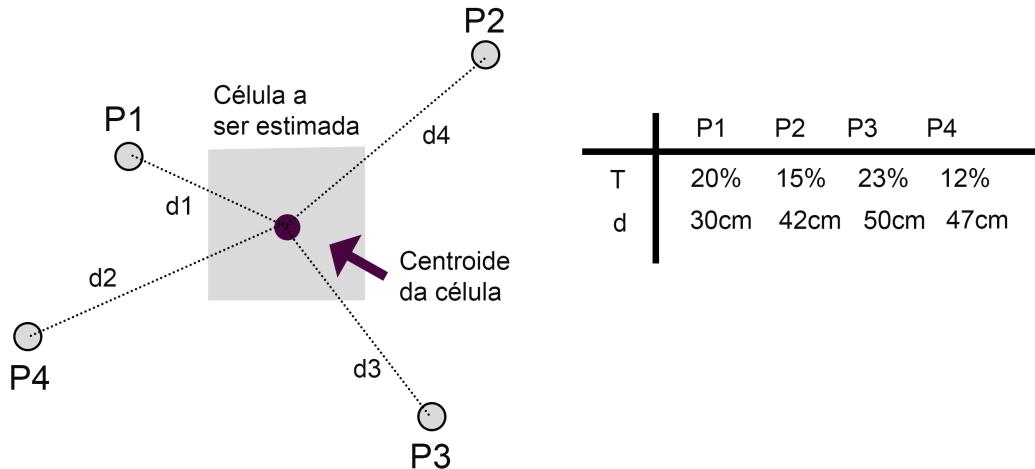


Figura 6.7: Obtenção dos volumes dos corpos geológicos a partir de extração de seções verticais interpretadas. Cinco seções consideradas e os respectivos furos de sondagens utilizados para interpretar os volumes.

Para 4 pontos mais próximos da célula estimada, temos respectivamente os valores de teor ( $T$ ) e a distância ao centroide da célula. Desta forma podemos calcular o valor médio da célula a partir de (6.8)

$$t_m = \frac{1/(30^2)20 + 1/(42^2)15 + 1/(50^2)23 + 1/(47^2)12}{1/(30^2) + 1/(42^2) + 1/(50^2) + 1/(47^2)} = 17,92\% \quad (6.8)$$

## 6.6 Tesselação de Delunay

A interpolação espacial pode ser realizada a partir da **tesselação de Delunay**, dividindo o espaço entre as amostras em triângulos. Cada amostra é univocamente ligada a dois pontos mais próximos, sendo esta uma solução geométrica única. Os valores médios de cada triângulo é obtido a partir da média ponderada entre os tamanhos da composta e os valores das propriedades. Dado três pontos  $\{P_1, P_2, P_3\}$  com respectivos valores de teor  $\{t_1, t_2, t_3\}$  e comprimentos  $\{l_1, l_2, l_3\}$ . O valor médio de cada triângulo pode ser obtido a partir de (6.9)

$$t_m = \frac{t_1l_1 + t_2l_2 + t_3l_3}{l_1 + l_2 + l_3} \quad (6.9)$$

Observe a figura 6.8. Estão dispostas sete amostras no espaço formando os

triângulos de Delunay, e o triângulo  $\{P_1, P_2, P_3\}$  está destacado.

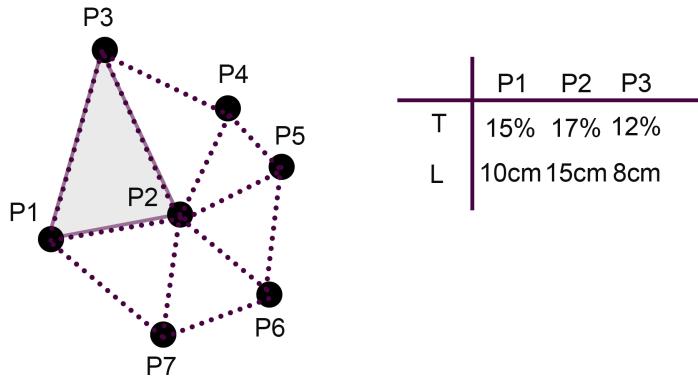


Figura 6.8: Valor médio obtido em um triângulo de Thiessen. 3 pontos amostrais  $\{P_1, P_2, P_3\}$ , com valores de teor  $\{t_1, t_2, t_3\}$ .

Considerando  $\{t_1, t_2, t_3\}$  os teores relativos a cada amostra, e  $\{l_1, l_2, l_3\}$  o tamanho das compostas. Podemos calcular o valor do teor médio como

$$t_m = \frac{15 * 10 + 17 * 15 + 12 * 8}{10 + 15 + 8} = 15,18\% \quad (6.10)$$

## 6.7 Polígonos de Thiessen

Uma das formas mais tradicionais de avaliação de propriedades georeferenciadas em duas dimensões é a utilização dos chamados polígonos de Thiessen. Cada polígono representa uma área correspondente de influência para uma determinada propriedade. Ao utilizarmos o princípio da generalização, podemos estender o valor de uma propriedade para toda a região considerada por este polígono. A disposição geométrica dos polígonos é única, todos eles são convexos e as relações espaciais destas figuras estão presentes em muitas questões envolvidas na natureza, como por exemplo, a formação de colméias ou de bolhas de sabão.

**Proposição 6.7.1** "Em 1911 um climatologista A. H. Thiessen sugeriu um método para representar a precipitação de dados baseados na disposição das estações de tempo. Ele definiu regiões baseadas em uma série de pontos no plano (estações de tempo) em "regiões mais próximas pela linha média entre as estações considerando as estações mais próximas". Baseado em sua proposta o termo polígono de Thiessen tem sido comumente utilizado na geografia definindo os polígonos formados pelo critério de proximidade no plano" -Brassel and Reif [1979]

Para criar um polígono de Tiessen é necessário realizar quatro etapas principais:

1. Determinar o ponto amostral considerado centroide do polígono. Unir aos pontos mais próximos semi-retas ligando o centroide.
2. Determinar os semi-planos formados pela reta perpendicular as semi-retas que ligam o centroide pela metade da distância entre eles.
3. Determinar os vértices do polígono a partir da interseção entre as retas determinadas no item 2.
4. Ligar todos os vértices do polígono. A solução é única e gerará um polígono convexo.

A figura 6.9 exemplifica a formação dos polígonos de Tiessen a partir da configuração geométrica de 6 pontos amostrais  $\{P_1, P_2, P_3, P_4, P_5, P_6\}$ , sendo  $P_1$  considerado o centroide do polígono. Qualquer propriedade tomada deste ponto amostral é estendida para toda a área do polígono considerado.

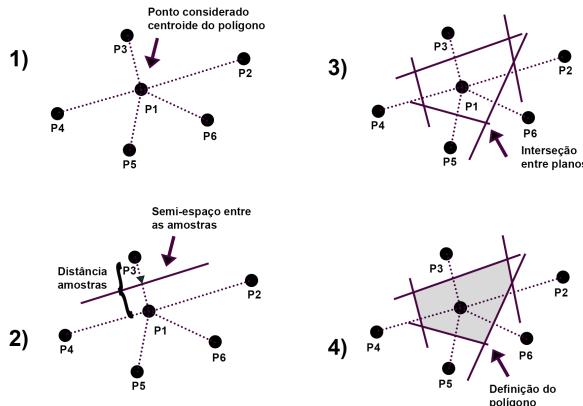


Figura 6.9: 4 etapas para a geração do polígono de Tiessen a partir de um ponto amostral considerado centroide ( $P_1$ ). 1) Determinar o ponto amostral considerado centroide do polígono. Unir aos pontos mais próximos semi-retas ligando o centroide. 2) Determinar os semi-planos formados pela reta perpendicular as semi-retas que ligam o centroide pela metade da distância entre eles. 3) Determinar os vértices do polígono a partir da interseção entre as retas determinadas no item 2. 4) Ligar todos os vértices do polígono. A solução é única e gerará um polígono convexo.

**Proposição 6.7.2** *A solução por polígonos de Tiessen é puramente geométrica. Nenhuma consideração é feita sobre a relação entre as propriedades de uma amostra no espaço. Apenas é realizada a extensão desta propriedade para uma área considerada de influência da amostra.*

Os polígonos de Thiessen possuem algumas propriedade geométricas. O vértice de um polígono de Thiessen é correspondente ao centro geométrico do círculo formado pelo centroide do polígono e dois pontos mais próximos. A figura 6.10 apresenta esta propriedade. Ao unir triângulos entre os pontos mais próximos é criada a chamada **tesselação de Delunay**, em que o baricentro do triângulo é também correspondente ao vértice do polígono de Thiessen.

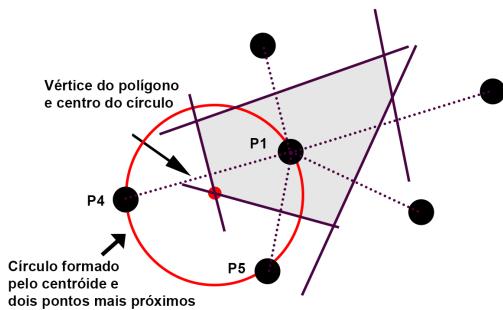


Figura 6.10: Propriedade dos polígonos de Thiessen. Centro do círculo composto pelo baricentro e dois pontos mais próximos forma um vértice do polígono de Thiessen.

É comum para os pontos que se situam nas extremidades do conjunto de dados não possuirem uma solução de polígono fechado, como ocorre nos pontos amostrais interiores. Neste caso geralmente se forma uma extração da área de influência. Esta extração pode considerar quesitos geológicos ou puramente geométricos, mas é um critério muito mais subjetivo que indicativo. A figura 6.12 apresenta esta extração.

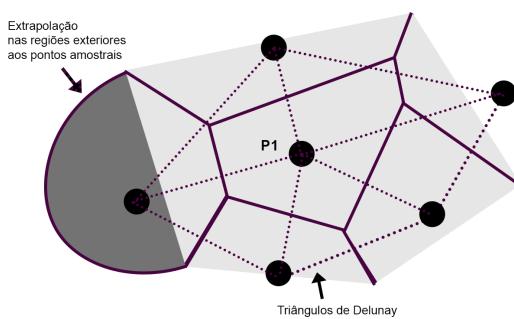


Figura 6.11: Extração realizada nos pontos amostrais situados na extremidade. Extração apresentada pelo um arco de cor cinza escuro.

Para obter o teor médio do depósito a partir dos polígonos de Thiessen basta

considerar a média ponderada entre as áreas de influência, tal como na equação

$$t_m = \frac{\sum_{i=1}^n A_i t_i}{\sum_{i=1}^n A_i} \quad (6.11)$$

Em que  $A$  representa a área do polígono de Thiessen para cada ponto amostral  $i$  e  $t$  representa o teor da área representado pelo valor da amostra em seu centroide. A solução tridimensional para os polígonos de Delunay é os chamados poliedros de Delunay. A solução tridimensional é complicada, e geralmente envolve elementos de topologia de alto nível computacional e matemático. A simplificação utilizada é utilizar o princípio do vizinho mais próximo, dividindo o espaço em uma malha de tamanho infinitesimal e atribuindo em cada célula a propriedade do ponto amostral mais próximo. Quando o tamanho da célula tende a um valor infinitesimal a solução pelo vizinho mais próximo converge para os polígonos ou poliedros de Thiessen. A figura 6.12 representa um mapa dos polígonos de Thiessen a partir do método do vizinho mais próximo e uma malha de tamanho unitário.

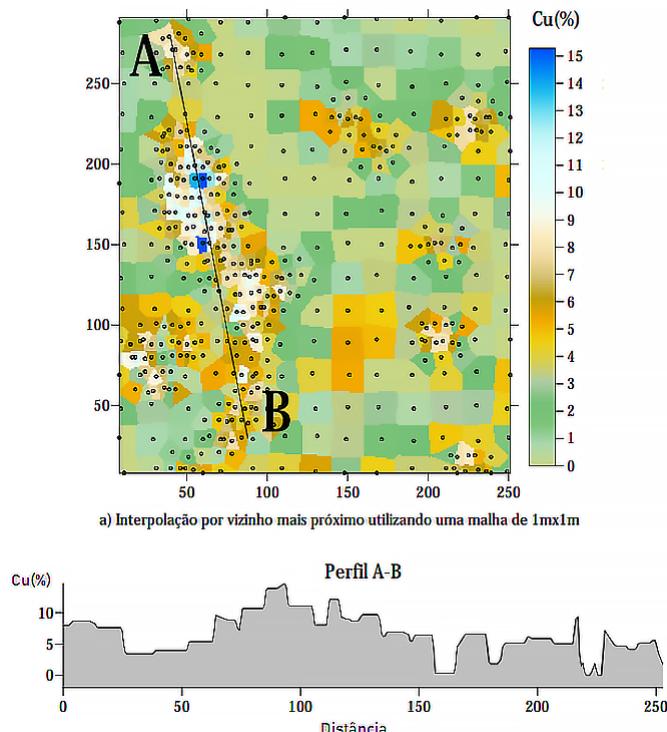


Figura 6.12: Aproximação dos polígonos de Thiessen a partir do vizinho mais próximo, utilizando uma malha de 1mx1m

## 6.8 Estatísticas desagrupadas

As malhas de amostragem representam uma importante fonte de informação para a realização dos métodos geoestatísticos. O posicionamento das amostras caracteriza a informação espacial a ser representada pelos métodos de estimativa. Quando pensamos em termos de representatividade a amostragem regular no espaço é a que melhor representa as características do depósito mineral. Em alguns casos é possível distribuir a malha segundo a continuidade espacial do depósito, permitindo um maior afastamento nas regiões mais contínuas do depósito e menos espaçadas nas regiões menos contínuas. A figura 6.13[A] apresenta esquematicamente uma malha regular e irregular. Em 6.13[B] é apresentado a disposição da malha segundo a continuidade do corpo geológico.

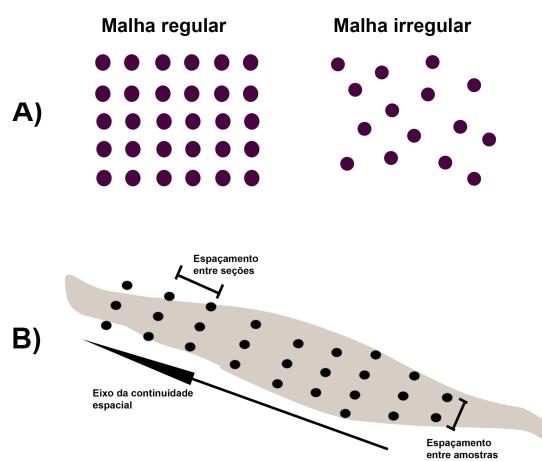


Figura 6.13: A) Representação de uma malha de amostragem regular e uma malha de amostragem irregular B) disposição de amostras segundo a continuidade espacial do corpo geológico.

Nos problemas de engenharia geológica, a disposição das malhas de sondagem dependem de diversos fatores, como por exemplo, terrenos de maior declividade que impedem a utilização de sondas, áreas de proteção ambiental, regiões de córregos ou outros fatores que podem impedir a formação de uma malha regular. Além disso, pela natureza de risco da atividade econômica, é comum adensar amostragens em lugares específicos onde possuam maior interesse para a mineração, como teores metálicos mais altos, ou litotipos de maior importância. Isto forma um agrupamento ou *cluster*. A figura 6.14 apresenta esquematicamente a formação de um agrupamento nas amostras.

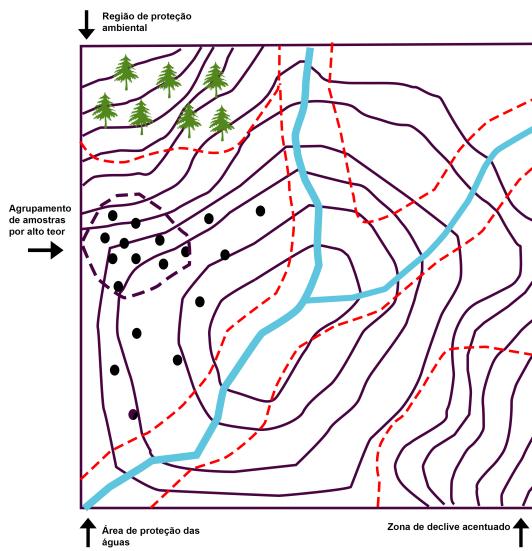


Figura 6.14: Representação de uma área amostrada. Obstáculos para a amostragem representados pela presença de áreas de preservação, terreno com maiores inclinações e área de reservas hídricas. Amostragem irregular realizada a oeste do desenho do mapa.

Calcular estatísticas considerando apenas os dados sem sua disposição espacial pode resultar em enviesamento. Se muitas análises são realizadas apenas em locais onde ocorre alto teor, os resumos estatísticos produziram também resultados com alto valor, mesmo que eles não correspondam à representação do domínio de estimativa.

**R** "É natural que os dados georeferenciados coletados são de uma forma não representativos. Amostragens preferenciais em áreas de interesse são intencionais e facilitadas pela intuição geológica, dados análogos e amostragens anteriores. A prática de coletar amostras agrupadas ou espacialmente enviesadas é encorajada pelas restrições técnicas e econômicas, tal como produções futuras, acessibilidade e custos de análise dos laboratórios" -[Pyrcz and Deutsch \[2003\]](#)

Um dos maiores erros cometidos por iniciantes ao considerar o desagrupamento de amostras é substituir os valores das amostras pelos valores dos pesos de desagrupamento. A alteração realizada pelo desagrupamento deve ser feita apenas sobre as estatísticas e não sobre seu valor bruto.

**Definição 6.8.1 — Desagrupamento.** *Dada uma estatística  $\phi(Z)$  a partir de uma variável aleatória  $Z$ , uma estatística desagrupada  $\theta$  é aquela que pode ser aplicada de tal forma que  $\theta(\phi(Z))$  considerando as distâncias euclidianas relativas entre as amostras.*

As duas principais técnicas utilizadas para desagrupamento são os polígonos de influência, ou de Thiessen vistos anteriormente e o desagrupamento por células.

### 6.8.1 Polígonos de influência

O desagrupamento das amostras pode ser realizado a partir de áreas de influência como no caso dos polígonos de Thiessen. A frequência de cada valor pode ser alterada pela área do polígono respectivamente. Observe a figura 6.15. Cada ponto amostral  $\{P_1, P_2, P_3, P_4, P_5, P_6\}$  possui uma área gerada pelo vizinho mais próximo em um grid de tamanho de célula conhecida. Abaixo podemos ver um histograma representando a frequência destes pontos. Se considerarmos apenas seus valores brutos, e não sua disposição espacial, cada ponto assume um valor de frequência igual a 1. No caso da utilização das áreas pelos polígonos

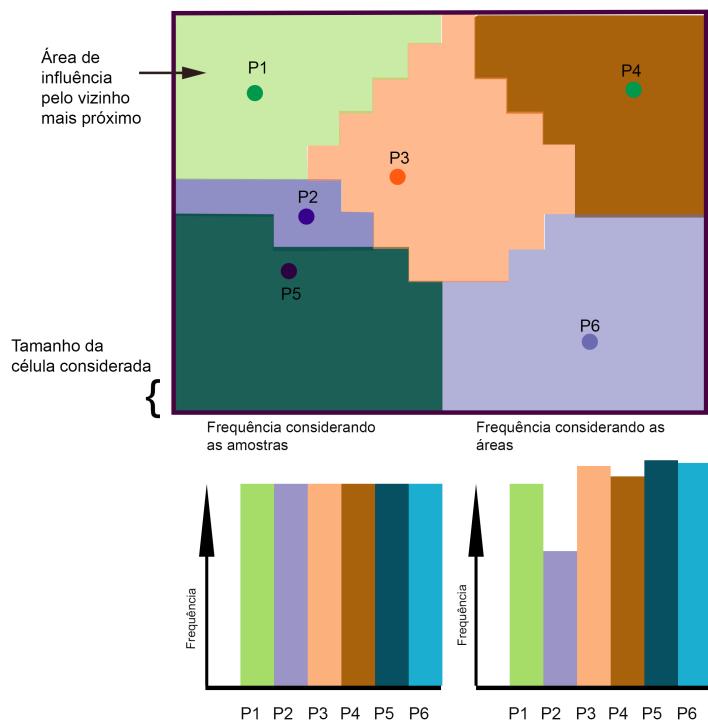


Figura 6.15: Representação de uma área amostrada. Obstáculos para a amostragem representados pela presença de áreas de preservação, terreno com maiores inclinações e área de reservas hídricas. Amostragem irregular realizada a oeste do desenho do mapa.

**Definição 6.8.2 — Desagrupamento por polígonos de influência.** *Dado uma amostra  $Z$  com uma realização  $z$ ,  $F(Z = z)$  representa a frequência de um elemento da amostra. Logo  $F(Z = z) = A(z)$ , sendo  $A(z)$  a área de influência de um elemento  $z$  da amostra.*

### 6.8.2 Desagrupamento por células

O desagrupamento realizado por polígonos de influência, gera uma solução única, e não permite encontrar pesos diferentes para as amostras, no entanto, o método de desagrupamento por células é flexível, permitindo ajustar parâmetros que indicarão o melhor resultado. O método considera a divisão do espaço em 'células' de mesma dimensão, tal que o peso de cada amostra é dado pelo número de amostras contidas dentro de cada célula. Observe a figura 6.16. A célula da linha 1 e coluna 1 apresenta apenas duas amostras, o que significa que cada uma receberá um peso de  $1/2$ .

	Coluna 1	Coluna 2	Coluna 3	Coluna 4	Coluna 5	Coluna 6
Linha 1	P1 P2	P3	P4 P5	P6		
Linha 2			P7 P8	P9		
Linha 3			P10	P11 P12	P13	
Linha 4				P14	P15	P16 P17

Figura 6.16: Representação do desagrupamento das células em um espaço bidimensional.

Evidentemente o tamanho da célula definirá o peso do desagrupamento. Uma célula muito grande que ocupe toda a extensão territorial analisada terá peso idêntico a  $1/n$ , sendo  $n$  o número de amostras. Logo o ponderador das amostras será igual a equação 6.12

$$p_{t_i} = (1/n) / \left( \sum_{i=1}^n 1/n \right) = 1/n \quad (6.12)$$

Exatamente igual a média aritmética dos valores. Da mesma forma se forem escolhidos tamanhos de células tão pequenas que apenas uma amostra esteja contida, teremos um valor de peso igual a 1, também obtendo o valor da média aritmética. A escolha do tamanho da célula deve ser feita entre estes dois casos extremos, aos quais teremos o menor valor desagrupado da média. A figura 6.17 demonstra a procura do tamanho da célula quadrada mais próxima do menor valor da média desagrupada, definindo assim o resultado que pretendemos.

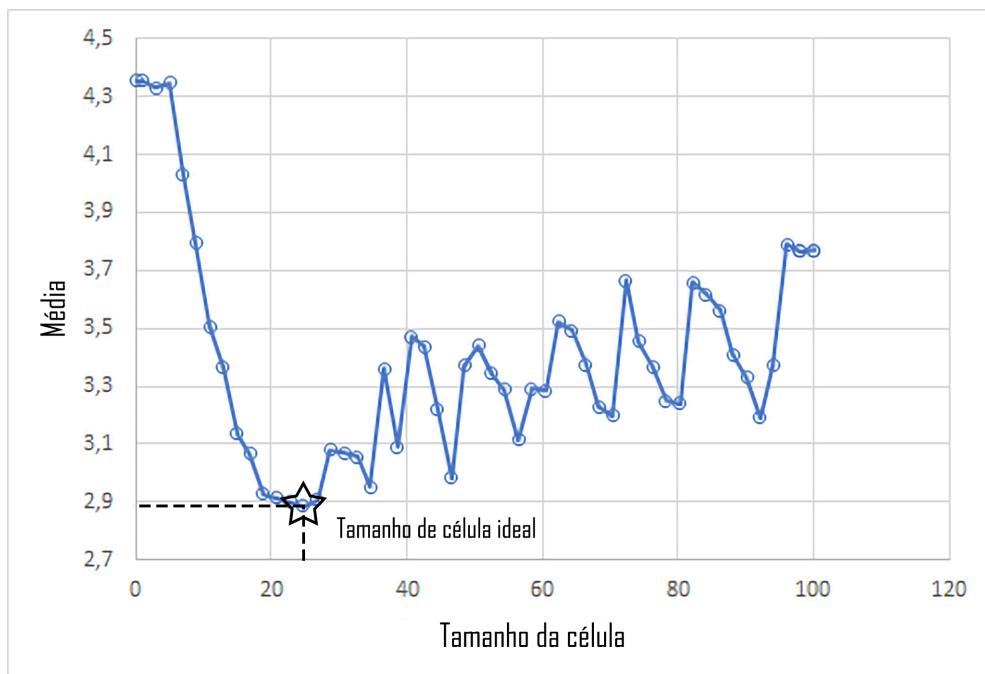
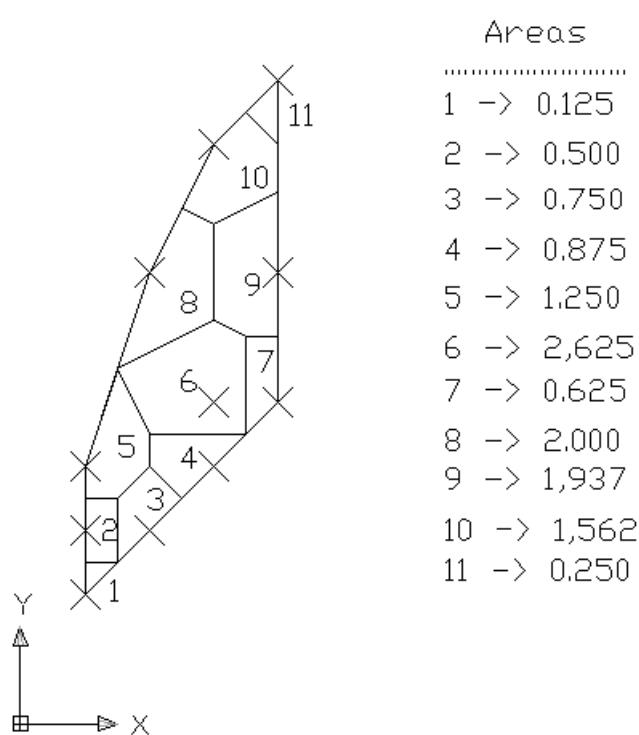


Figura 6.17: Representação da escolha do melhor tamanho de célula dado um conjunto de médias desagrupadas.

**Exercícios 6.1** Os dados da tabela abaixo representam um conjunto de amostras bidimensionais, em que  $x$  e  $y$  representam respectivamente as coordenadas cartesianas nos eixos das abscissas 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



**Exercícios 6.2** Para os dados do exercício anterior encontre a média declusterizada e a variância declusterizada dos dados.



## 7. Continuidade Espacial

*Uma arte que tem vida não reproduz o passado; ela dá continuidade a ele.*

*Auguste Rodin*

Nos capítulos anteriores aprendemos como a função aleatória é uma função do domínio considerado e que valores de uma propriedade proximais tendem a ser mais parecidos entre si, introduzindo o conceito de **continuidade**. Neste capítulo introduziremos a **análise de continuidade espacial** demonstrando as ferramentas necessárias para entendermos o quanto dependente variáveis aleatórias se encontram no espaço. Modelos de continuidade espacial são o fundamento para a análise estatística de muitos modelos, como os modelos de estimativa e simulação. No caso de modelos de estimativa como as krigagem ordinária e simples, que veremos nos próximos capítulos, a continuidade espacial influencia menos nos resultados se comparado aos métodos de simulação.

### 7.1 Definição de continuidade espacial e variografia

A continuidade espacial é uma relação do grau de dependência de valores no espaço. Podemos pensar na continuidade a partir de uma analogia com um terreno. Um terreno mais suave significa que a diferença de valores mais próximos entre si, e por isso apresenta maior **continuidade**. Um terreno mais grosseiro, no entanto,

é menos contínuo, apresentando uma rugosidade maior. A figura 7.1 apresenta as diferenças entre um fenômeno mais contínuo de um menos contínuo.

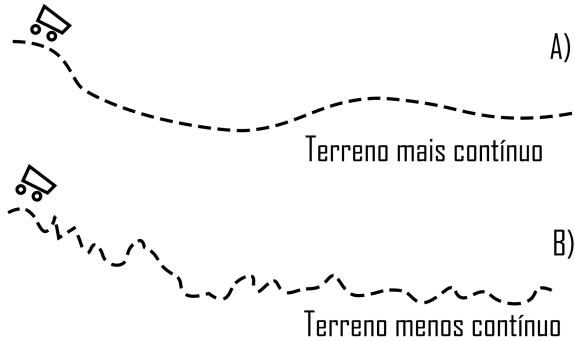


Figura 7.1: Analogia da continuidade espacial com a rugosidade de um terreno. Em A) temos um terreno mais contínuo, com menor rugosidade. Em B) temos um terreno menos contínuo, ou mais rugoso.

A ideia de continuidade espacial também está associada a fractais e de auto-similaridade. Um fenômeno auto-similar é aquele capaz de reproduzir características parecidas em diferentes escalas apresentadas. A auto-similaridade é algo presente na geologia de depósitos minerais pois os corpos geológicos tendem a apresentar características muito parecidas quando consideramos desde a escala litológica até os macrodomínios geológicos. Na geoestatística, as medidas de continuidade são avaliadas em formulações bi-pontuais considerando a disposição de variáveis deslocadas a partir de um vetor  $h$ , dentro de um domínio  $D$ . A figura 7.2 demonstra a relação entre duas variáveis  $Z(u_1)$  e  $Z(u_1 + h)$  considerando um vetor  $h$ .

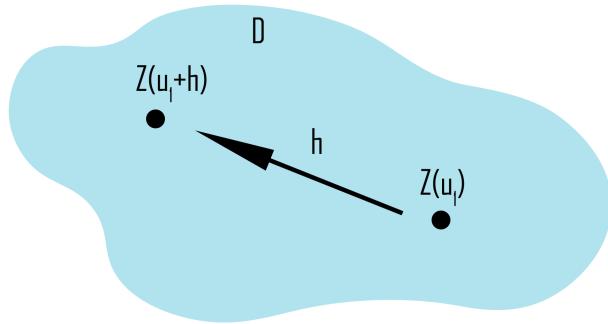


Figura 7.2: Notação geoestatística para a definição de um lag  $h$  em uma direção dentro de um domínio  $D$  de uma amostra com suporte  $x$ . Dois pares de pontos, um considerado head value, ou ponta do vetor e outro considerado tail value, ou início do vetor.

A premissa da determinação da continuidade espacial é que ela é **invariável por translação**. Dada uma direção de um vetor  $h$ , qualquer dependência entre as

variáveis pode ser unicamente expressa pelo comprimento e direção do vetor, sendo as diferenças apresentadas neste *lag* valores auto-similares.

**Proposição 7.1.1** *Muitas das leis físicas partem do pressuposto vetorial, pois as leis tendem a ser reproduutíveis no espaço independente de sua posição, mas considerando suas posições relativas. A continuidade espacial é uma medida do fenômeno reproduzido pela função aleatória  $Z(u)$ , e não por valores individuais dispostos no espaço. Este é um erro muito cometido por inciantes da estatística que acreditam que a continuidade espacial depende das amostras, e na verdade são reproduzidas por uma lei geradora. Pelo fato de não conhecermos as disposições amostrais em nível extensivo, o conhecimento da lei de continuidade espacial é blindado, dado uma falsa impressão que esta é dependente dos valores amostrais.*

As funções de continuidade espacial medem comportamentos diferentes da variável aleatória: ou elas determinam a **similaridade entre variáveis aleatórias**, ou determinam suas **dissimilaridades**. Em termos das diferenças entre as variáveis, podemos expressar o **variograma** pela equação 7.1

$$2\gamma(h) = E \{ [Z(u) - Z(u + h)]^2 \} \quad (7.1)$$

Em que  $2\gamma(h)$  é chamado de variograma e  $\gamma(h)$  é chamado de semi-variograma. Na literatura é muito comum ocorrem diferenças de nomenclatura e o semi-variograma pode ser chamado de variograma. A diferença, no entanto, é puramente uma modificação de escala analisada, e por isso quando utilizados softwares comerciais ou gratuitos, é importante ao aluno verificar com cuidado o tipo de função utilizada. Da mesma forma a função **covariograma** representa o grau de similaridade entre variáveis tal como a equação 7.2

$$C(h) = E \{ [Z(u) * Z(u + h)] \} - E(Z(u))E(Z(u + h)) \quad (7.2)$$

Quando considerada a hipótese de estacionariedade de segunda ordem, os valores médios de  $Z(u)$  e  $Z(u+h)$  são idênticos a  $m$ . Logo a equação 7.2 pode ser substituída por 7.3

$$C(h) = E \{ [Z(u) * Z(u + h)] \} - m^2 \quad (7.3)$$

Podemos relacionar os valores de similaridade e dissimilaridade das variáveis

aleatórias se considerarmos a hipótese de estacionaridade de segunda ordem. Neste caso definimos

$$\gamma(h) = C(0) - C(h) \quad (7.4)$$

Em que  $C(0)$  chamado de **variância a priori** é igual a  $Var(Z(u)) = Var(Z(u + h))$ , dadas as condições de estacionaridade de segunda ordem. Considerando os vetores amostrais  $Z(u)$  e  $Z(u + h)$ , as funções covariograma e variograma podem ser equiparadas a projeções vetoriais, onde  $C(h)$  é análogo ao produto escalar entre estes vetores, e  $\gamma(h)$  pode ser representado como uma diferença de vetores. As funções covariograma e variograma são obtidas a partir do conhecimento da função aleatória  $Z(u)$ , o que em prática não conhecido. O objetivo principal é obter uma estatística  $\hat{\gamma}(h)$  que aproxime da função  $\gamma(h)$  a partir do conhecimento dos dados. Os cálculos a partir dos dados são denominados de **funções experimentais**, enquanto a determinação de um modelo  $\hat{\gamma}(h)$  é chamado de **modelagem da continuidade**.

## 7.2 Procura de pares de amostras no espaço

O conhecimento da função aleatória  $Z(u)$ , como dito anteriormente, é impossível de ser encontrado. Para isso é necessário conhecer  $Z(u)$  e  $Z(u + h)$  a partir de pares conjugados de amostras  $z(u)$  e  $z(u + h)$ . Dado um vetor  $h$  de direção específica e tamanho correspondente a  $|h|$ , podemos encontrar os pares de variáveis de acordo com a figura 7.3. A cada deslocamento do vetor  $h$  na direção considerada obtemos duas amostras, uma na cabeça do vetor e outra no início. Estes pares podem ser plotados em um gráfico de dispersão como visto no capítulo de estatística bivariada, assim construindo uma nuvem de dispersão segundo aquela direção.

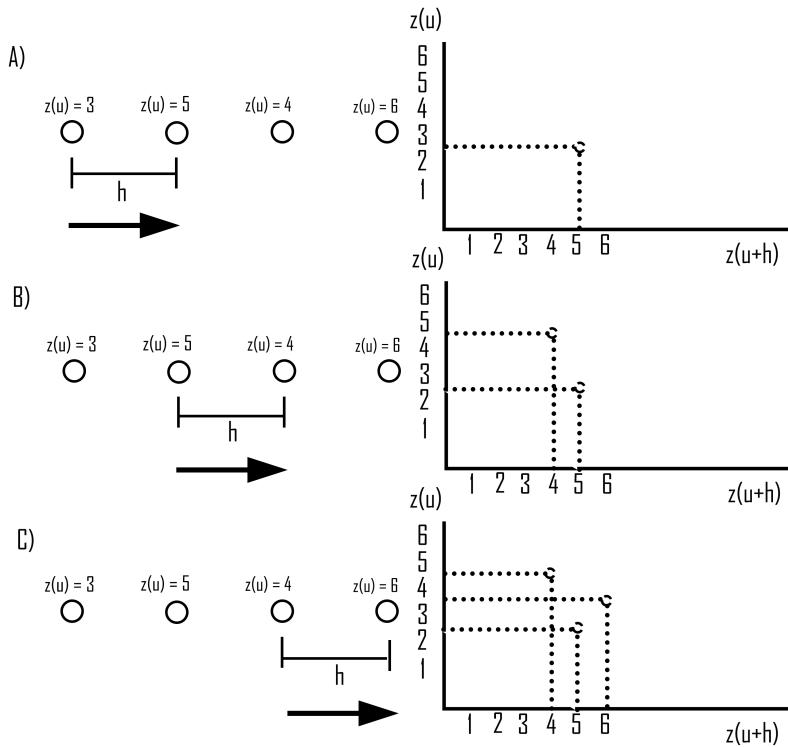


Figura 7.3: Determinação de pares de amostras segundo um vetor de distância  $h$ .  
A) Primeiro deslocamento do vetor. B) Segundo deslocamento do vetor. C) Terceiro deslocamento do vetor.

Quando aumentamos o valor do tamanho do lag  $h$ , os pares conjugados serão outros. Podemos observar na figura 7.4 a determinação do novo gráfico de pontos.

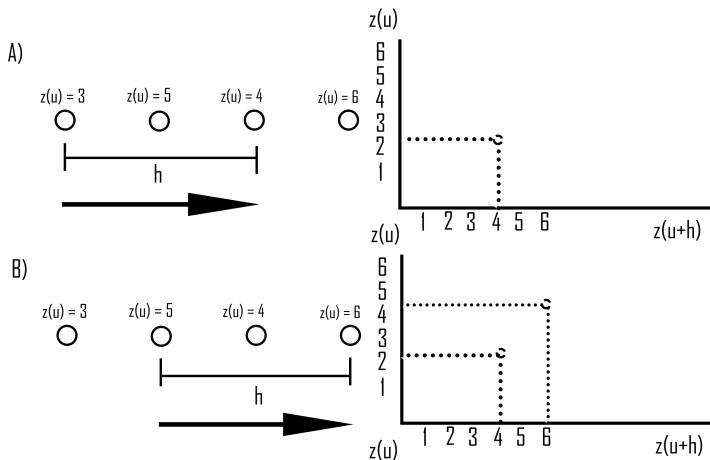


Figura 7.4: Determinação de pares de amostras segundo um vetor de distância  $h$ .  
A) Primeiro deslocamento do vetor. B) Segundo deslocamento do vetor. C) Terceiro deslocamento do vetor.

Para os fenômenos geológicos, é de se esperar que as amostras mais próximas

apresentem maior similaridade de valores amostrados. A Figura 7.5 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.

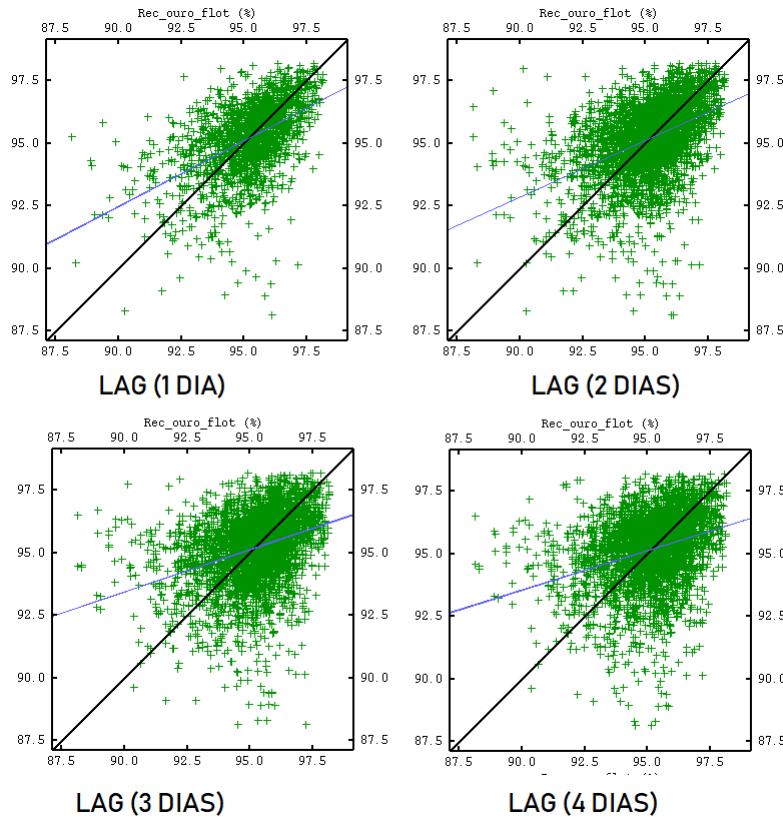


Figura 7.5: H-scatterplots de uma variável para lags de 0,1m , 0,2m, 0,3m e 0,4m. Aumento da descorrelação de acordo com o aumento do comprimento dos lags.

## 7.3 Funções experimentais de continuidade espacial

### 7.3.1 Efeito dos dados sobre os valores experimentais

As funções clássicas de continuidade espacial são afetadas por valores extremos, esparcidez 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.

### 7.3.2 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 7.5:

$$C(h) = E[(Z(x + h) - m(x + h))(Z(x) - m(x))] \quad (7.5)$$

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 estacionariedade de segunda ordem os valores da média são constantes tal que  $m(x + h) = m_i$  e a Equação 7.5 pode ser traduzida pela Equação 7.6 por uma transformação algébrica:

$$C(h) = E(Z(x + h)Z(x)) - m^2 \quad (7.6)$$

A função variograma pode ser representada pela Equação 7.7:

$$2\gamma(h) = E [(Z(x + h) - Z(x))^2] \quad (7.7)$$

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 7.6 é uma demonstração gráfica da interpretação do semivariograma como um valor médio das distâncias das variáveis  $Z_u$  e  $Z_{u+h}$ , com esses valores em x e y e com a reta de correlação máxima. A demonstração da Figura 7.6 em termos matemáticos está descrito na Equação 7.8:

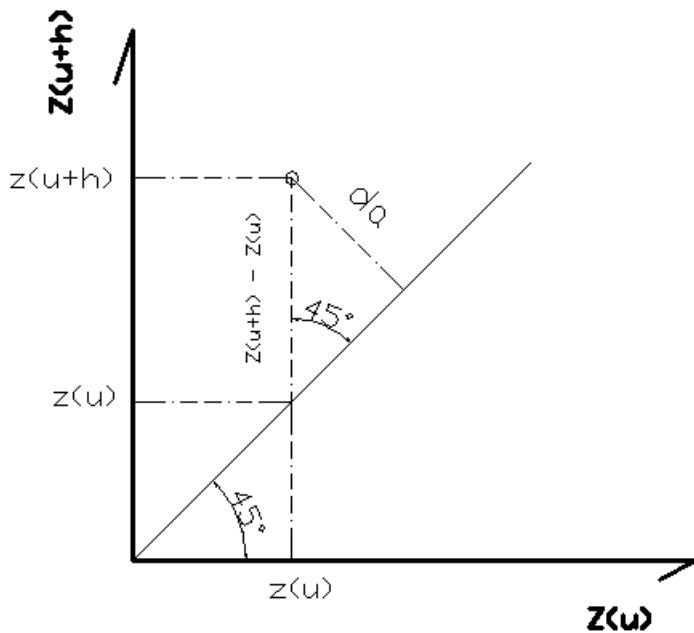


Figura 7.6: Interpretação geométrica do semi-variograma como sendo a distância de um ponto em relação a reta  $x=y$ .

$$\frac{1}{n} \sum_{i=1}^n da^2 = \frac{1}{n} \sum_{i=1}^n \operatorname{sen}^2(45^\circ) (z(x) - z(x + h))^2 = \frac{1}{2n} \sum_{i=1}^n (z(x) - z(x + h))^2 \quad (7.8)$$

### 7.3.3 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 7.9 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:

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^n |Z_i - Z_{i+h}|^{\frac{1}{2}} \quad (7.9)$$

A Tabela 7.1 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 7.1: Funções de continuidade espaciais experimentais

Função experimental	Equação
Semi-variograma	$\sum_{i=0}^n \frac{(Z_i - Z_{i+h})^2}{2n}$
Covariograma	$\frac{1}{n} \sum_{i=0}^n (Z_i - m_i) \cdot (Z_{i+h} - m_{i+h})$
Correlograma	$\frac{1}{n} \sum_{i=0}^n \frac{(Z_i - m_i) \cdot (Z_{i+h} - m_{i+h})}{\sigma_i \sigma_{i+h}}$
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  Z_i - Z_{i+h} $
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 7.7 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.

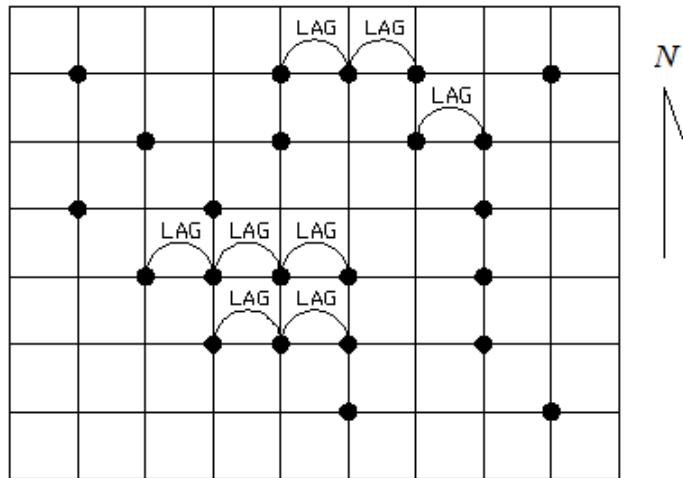


Figura 7.7: Cálculo de variogramas experimentais segundo um lag unitário na direção Leste-Oeste.

O mesmo pode ser representado na Figura 7.8, em que os valores disponíveis como pares para o cálculo são efetuados em dois nós do grid consecutivos.

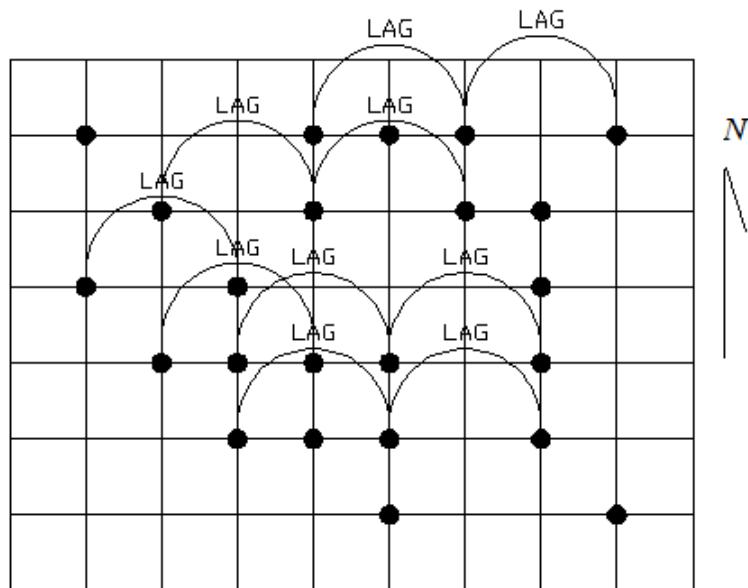


Figura 7.8: Cálculo de variogramas experimentais segundo o dobro do lag na direção leste-oeste.

Os valores calculados do variograma nas Figuras 7.7 e 7.8 estão demonstrados em um gráfico na Figura 7.9 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.

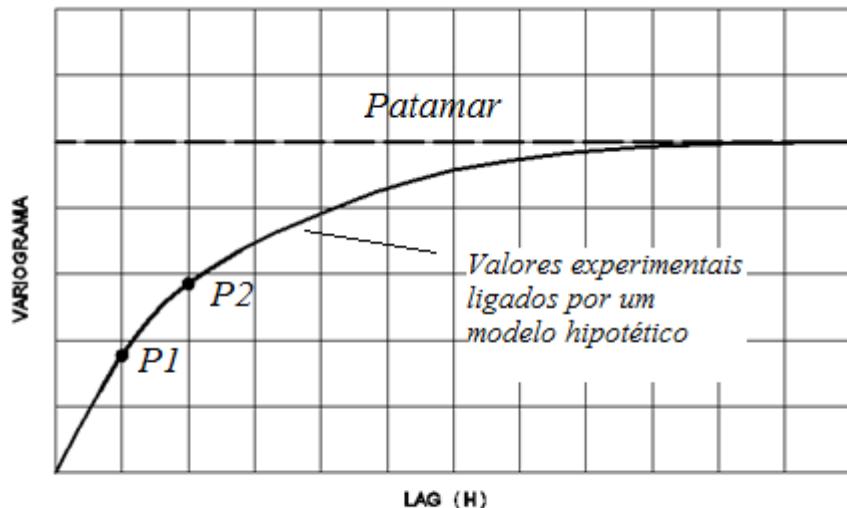


Figura 7.9: Função variograma experimental para os cálculos nas Figuras 9 e 10 e ajuste em um modelo hipotético. P1 e P2 representam os pontos para um lag unitário e o dobro do lag.

#### 7.3.4 Parâmetros de busca

Nas Figuras 7.7 e 7.8, 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 7.10 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.

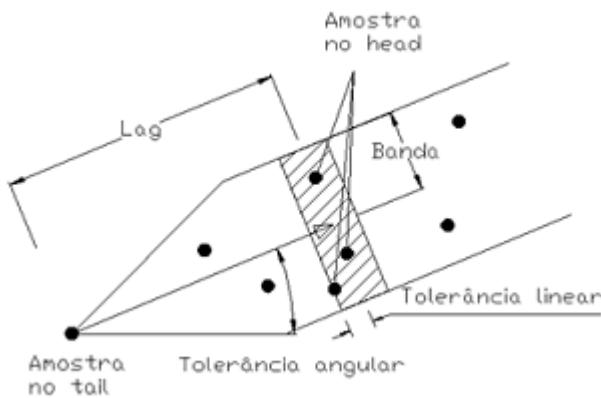


Figura 7.10: Busca de pontos da função de continuidade para amostras irregularmente espaçadas.

No caso tridimensional, a busca de pares pode se realizar de duas formas diferentes, sob uma perspectiva elíptica ou prismática. A Figura 7.11 é uma representação da busca de pares nas duas formas geométricas.

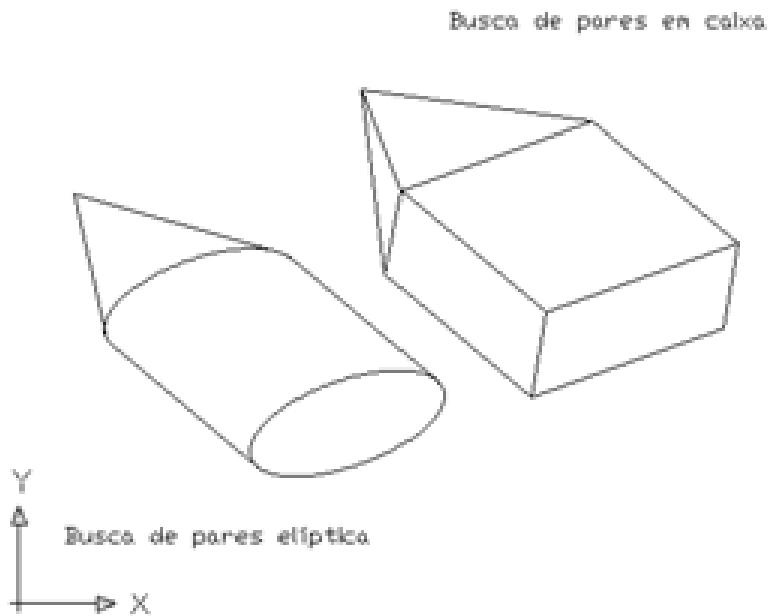


Figura 7.11: Busca de pares de pontos em uma direção tridimensional. Busca de pares elíptica utilizada no SGeMS e em caixa utilizada no GSLib.

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.

## 7.4 Modelagem de funções de continuidade espacial

### 7.4.1 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 7.10.

$$\sum_{i=0}^n \sum_{j=0}^n \lambda_i \lambda_j \gamma(x_i - x_j) \geq 0 \quad (7.10)$$

Em que  $\lambda_i$  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.

### 7.4.2 Parâmetros das funções de continuidade

O conjunto de variogramas transitivos, ou seja, que apresentam um patamar possuem parâmetros característicos. A Figura 7.12 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 7.13. 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.

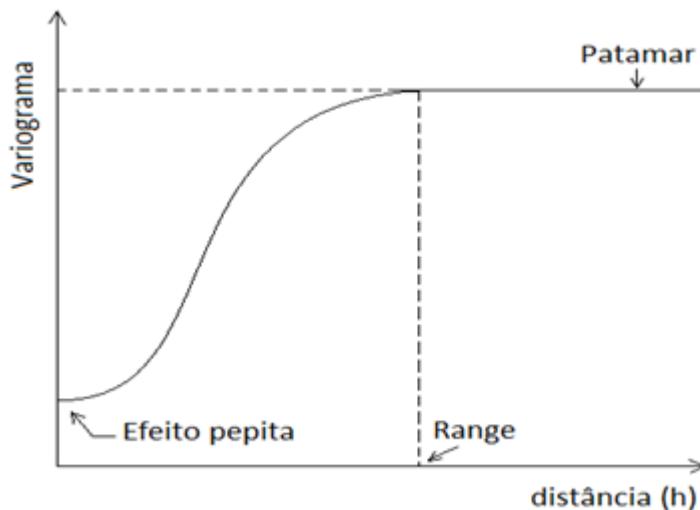


Figura 7.12: Parâmetros do variograma.

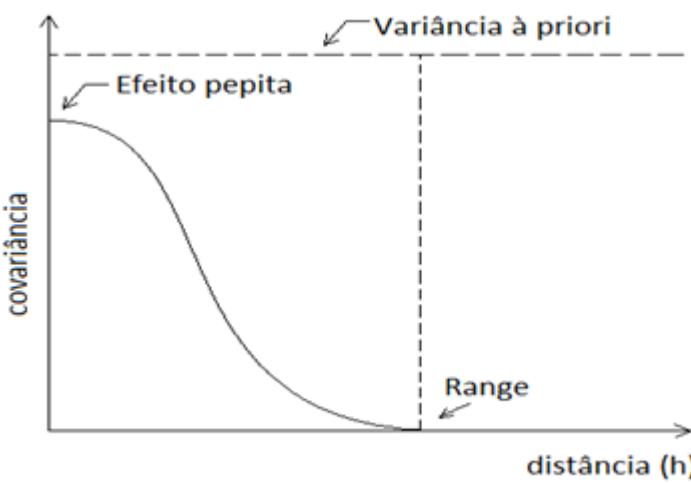


Figura 7.13: Parâmetros da função covariograma.

#### 7.4.3 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 7.11 demonstra o modelo de variograma com efeito de pepita puro.

$$\gamma(h) = \begin{cases} 0 & , h = 0 \\ 1 & , \text{ao contrário} \end{cases} \quad (7.11)$$

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 7.12 demonstra o modelo de variograma exponencial.

$$\gamma(h) = 1 - \exp^{\frac{-h}{a}} \quad (7.12)$$

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 7.13 demonstra o modelo de variograma gaussiano.

$$\gamma(h) = 1 - \exp^{-\frac{h^2}{a^2}} \quad (7.13)$$

Modelo Esférico: A Equação 7.14 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.

$$\gamma(h) = \begin{cases} \left( \frac{3h}{2a} - \frac{h^3}{2a^3} \right) & , h < a \\ 1 & , h \geq a \end{cases} \quad (7.14)$$

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

#### 7.4.4 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 7.14 é uma representação da anisotropia geométrica.

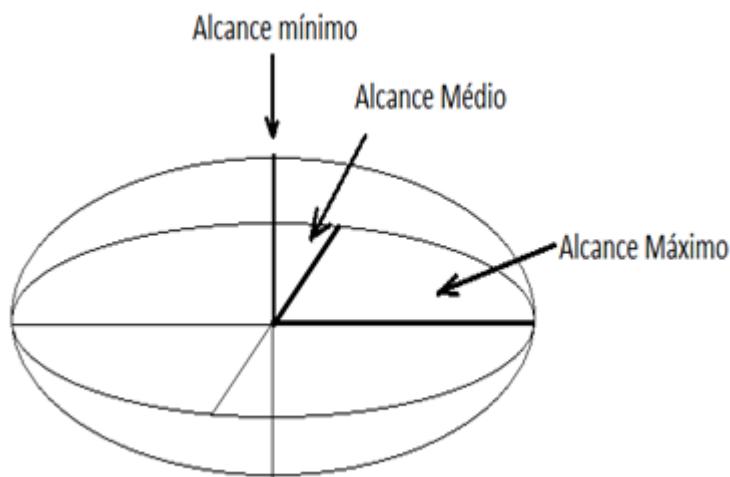


Figura 7.14: Representação do modelo de anisotropia geométrico. Elipsóide com valores de alcance mínimo, médio e alcance máximo.

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 7.15 representa a matriz de rotação das coordenadas para os eixos de referência.

$$Q = \begin{bmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \cos\theta_2 & \sin\theta_2 & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 \end{bmatrix} \begin{bmatrix} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (7.15)$$

Em que  $\theta_3$  consiste no ângulo de rotação no eixo z,  $\theta_2$  o ângulo de rotação no eixo y e  $\theta_1$  a rotação do ângulo no eixo x. Os valores do vetor unitário são então

redimensionados segundo a matriz de dilatação da Equação 7.16.

$$D = \begin{bmatrix} l_1 & 0 & 0 \\ 0 & l_2 & 0 \\ 0 & 0 & l_3 \end{bmatrix} \quad (7.16)$$

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 7.15 que a diferença de azimute pelo ângulo  $\theta$  leva a uma diferença de patamares de  $g_1$  para  $g_1 + g_2$ . 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.

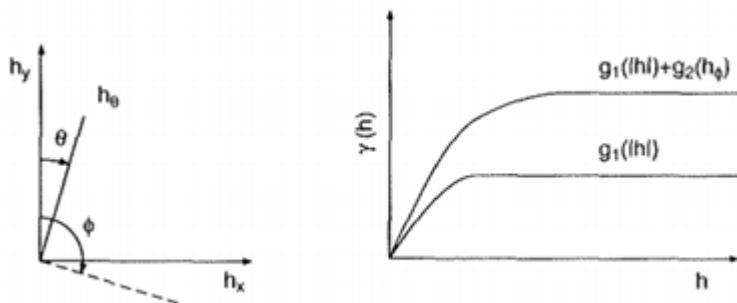


Figura 7.15: Anisotropia zonal representada nos variogramas a) Diferença entre as direções dos variogramas b) Representação da anisotropia por variogramas com patamares diferentes.

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.

### 7.4.5 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 7.17 em que  $Z_i$  e  $Z_j$  são variáveis distintas e  $m_i$  e  $m_j$  são suas respectivas médias:

$$C_{ij}(h) = \frac{1}{2}E[(Z_i(x+h) - m_i)(Z_j(x) - m_j)] \quad (7.17)$$

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 7.18:

$$C_{ij}(h) = \frac{1}{2}(C_{ij}(h) + C_{ij}(-h)) + \frac{1}{2}(C_{ij}(h) - C_{ij}(-h)) \quad (7.18)$$

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 7.19 expressa a fórmula da função:

$$\gamma_{ij}(h) = \frac{1}{2}E[(Z_i(x+h) - Z_i(x))(Z_j(x+h) - Z_j(x))] \quad (7.19)$$

#### 7.4.6 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.

#### 7.4.7 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 7.20 demonstra a relação de dissimilaridade entre o modelo e os variogramas experimentais:

$$\psi = \sum_{i=0}^n \rho_i (\gamma_i - \gamma_i^*) \quad (7.20)$$

Em que  $\psi$  é a equação objetivo,  $\gamma_i$  são os valores experimentais e  $\gamma_i^*$  são os valores de um modelo a ser ajustado, para uma função de ajuste  $\rho$ .

**Exercícios 7.1** 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

■



## 8. Krigagem

### 8.1 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 unbiased 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 (8.1)

$$z_0 = \sum_{i=1}^n \lambda_i z_i \quad (8.1)$$

Em que  $\lambda_i$  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 capítulo um podemos definir o erro de estimativa segundo a equação (8.2)

$$\varepsilon(z_0^*) = z_0^* - z_0 \quad (8.2)$$

Em que  $z_0^*$  é o valor estimado no ponto desconhecido e  $z_0$  é o valor real naquele ponto. Substituindo a equação (8.1) em (8.2) obtemos a relação (8.3)

$$\varepsilon(z_0^*) = \sum_{i=1}^n \lambda_i z_i - z_0 \quad (8.3)$$

Tomando o quadrado do erro de estimativa temos a seguinte a relação (8.4)

$$\varepsilon(z_0^*)^2 = \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j z_i z_j - 2 \sum_{i=1}^n \lambda_i z_i z_0 - z_0^2 \quad (8.4)$$

Tomando o menor valor esperado do erro quadrático temos então

$$E(\varepsilon(z_0^*)^2) = \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j Cov(z_i, z_j) - 2 \sum_{i=1}^n \lambda_i Cov(z_i, z_0) - Cov(z_0, z_0) \quad (8.5)$$

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 6 no subitem 6.8 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 (8.5)

$$\sigma_{krig}^2 = \frac{\partial}{\partial \lambda_i} \left( \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j Cov(z_i, z_j) - 2 \sum_{i=1}^n \lambda_i Cov(z_i, z_0) - Cov(z_0, z_0) \right) + \frac{\partial}{\partial \lambda_i} R = 0 \mid \forall i \quad (8.6)$$

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 (8.7)

$$\sigma_{krig}^2 = 2 \sum_{i=1}^n \lambda_i Cov(z_i, z_i) - 2Cov(z_i, z_0) + \frac{\partial}{\partial \lambda_i} R = 0 \mid \forall i \quad (8.7)$$

O cálculo das derivadas parciais pode ser realizado de acordo com a expansão dos somatórios como demonstrado na equação (8.5)

$$\sum_{i=1}^n \sum_{j=1}^n \lambda_j \lambda_i Cov(z_i, z_j) = \sum_{j=2}^n \lambda_j \lambda_1 Cov(z_1, z_j) + \sum_{i=2}^n \lambda_1 \lambda_i Cov(z_i, z_1) + \lambda_1 \lambda_1 Cov(z_1, z_1) \quad (8.8)$$

Cada sistema de krigagem pode então ser resolvido de acordo com uma matriz de covariâncias genérica como descrito em (8.6)

$$\begin{pmatrix} Cov(z_1, z_1) & Cov(z_1, z_2) & \dots & Cov(z_1, z_n) & \frac{\partial R}{\partial \lambda_1} \\ Cov(z_2, z_1) & Cov(z_2, z_2) & \dots & Cov(z_2, z_n) & \frac{\partial R}{\partial \lambda_2} \\ \dots & \dots & \dots & \dots & \dots \\ Cov(z_n, z_1) & Cov(z_n, z_2) & \dots & Cov(z_n, z_n) & \frac{\partial R}{\partial \lambda_n} \\ 1 & 1 & 1 & \dots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ 1/2 \end{pmatrix} = \begin{pmatrix} Cov(z_0, z_1) \\ Cov(z_0, z_2) \\ \dots \\ Cov(z_0, z_n) \\ P \end{pmatrix} \quad (8.9)$$

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 kriga-

gem 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 (8.2)

## 8.2 Krigagem Ordinária

Para a krigagem ordinária utilizamos os mesmos pressupostos e cálculos utilizados em 8.1. Logo temos como restrição à condição de não viés dado pela demonstração abaixo:

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

■

Logo nossa função de restrição pode ser determinada por (8.10), e o nosso valor P é igual a 1.

$$R_i = \mu \left( \sum_{i=0}^n \lambda_i - 1 \right) \quad | \forall i \tag{8.10}$$

Em que  $\mu$  é o multipliador lagragiano. Logo tomando a derivada parcial de cada uma das restrições para cada uma das amostras i do problema temos a relação segundo a equação (8.11)

$$\frac{\partial}{\partial \lambda_i} R_i = \mu \quad | \forall i \tag{8.11}$$

O sistema de equações da krigagem pode ser transformado então na relação 8.12

$$\begin{pmatrix} \text{Cov}(Z_1, Z_1) & \text{Cov}(Z_1, Z_2) & \dots & \text{Cov}(Z_1, Z_n) & 1 \\ \text{Cov}(Z_2, Z_1) & \text{Cov}(Z_2, Z_2) & \dots & \text{Cov}(Z_2, Z_n) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ \text{Cov}(Z_n, Z_1) & \text{Cov}(Z_n, Z_2) & \dots & \text{Cov}(Z_n, Z_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} \text{Cov}(Z_0, Z_1) \\ \text{Cov}(Z_0, Z_2) \\ \dots \\ \text{Cov}(Z_0, Z_n) \\ 1 \end{pmatrix} \quad (8.12)$$

### 8.3 Krigagem Simples

A krigagem simples tem como pressuposto encontrar os ponderadores que minimizem o resíduo da variável aleatória. Como determinado em ??, nada mais é que a própria variável subtraída 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 (8.13)

$$Z_0^* = \sum_{i=0}^n \lambda_i (Z_i - m) + m \quad (8.13)$$

Podemos isolar os termos da equação (8.7) em relação ao valor médio assim obtendo a equação (8.14)

$$Z_0^* = \sum_{i=0}^n \lambda_i Z_i + m \left( 1 - \sum_{i=0}^n \lambda_i \right) \quad (8.14)$$

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

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

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

$$\sum_{i=0}^n \lambda_i E(Y_0) + m = m$$

$$\sum_{i=0}^n \lambda_i E(Y_0) = 0$$

$$E(Y_0) = 0 \vee \sum_{i=0}^n \lambda_i = 0$$

■

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 (8.15) e o nosso valor P de restrição à soma dos ponderadores é igual a 0.

$$R_i = \mu \sum_{i=0}^n \lambda_i | \forall i \quad (8.15)$$

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

$$\frac{\partial}{\partial \lambda_i} R_i = \mu | \forall i \quad (8.16)$$

Logo o sistema de matrizes para a krigagem simples pode ser transformado em (8.17)

$$\begin{pmatrix} Cov(Y_1, Y_1) & Cov(Y_1, Y_2) & \dots & Cov(Y_1, Y_n) & 1 \\ Cov(Y_2, Y_1) & Cov(Y_2, Y_2) & \dots & Cov(Y_2, Y_n) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & \dots & Cov(Y_n, Y_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} Cov(Y_0, Y_1) \\ Cov(Y_0, Y_2) \\ \dots \\ Cov(Y_0, Y_n) \\ 0 \end{pmatrix} \quad (8.17)$$

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

## 8.4 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 tomado seu valor médio. A figura (8.1) demonstra um bloco B contendo novo pontos k1 até k9 contidos nele.

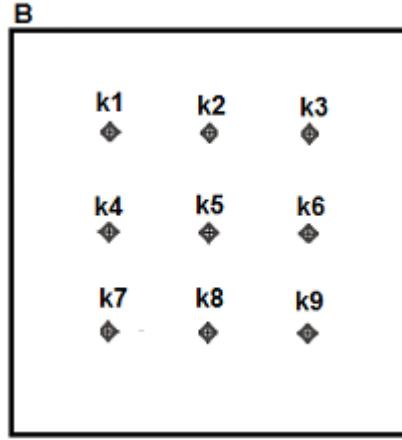


Figura 8.1: Demonstração dos pesos de krigagem para o posicionamento de amostras para um modelo de pepita puro

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 (8.18)

$$Z_v = \frac{1}{p} \sum_{j=1}^p Z_j \quad (8.18)$$

Em que  $p$  é o número de variáveis contidas dentro daquele volume. Logo temos o erro de estimativa dado pela equação (8.19)

$$\varepsilon(Z_v^*) = Z_v^* - Z_v \quad (8.19)$$

$$\text{Demonstração. } \varepsilon(Z_v^*) = \frac{1}{p} \sum_{j=1}^p (\sum_{i=1}^n \lambda_i Z_i - Z_j)$$

$$\varepsilon(Z_v^*) = \frac{1}{p} \sum_{j=1}^p \sum_{i=1}^n \lambda_i Z_i - \frac{1}{p} \sum_{j=1}^p Z_j$$

$$\varepsilon(Z_v^*) = \frac{1}{p} p \sum_{i=1}^n \lambda_i Z_i - \frac{1}{p} \sum_{j=1}^p Z_j$$

$$\varepsilon(Z_v^*) = \sum_{i=1}^n \lambda_i Z_i - \frac{1}{p} \sum_{j=1}^p Z_j$$

$$\varepsilon(Z_v^*)^2 = \sum_{i=1}^n \sum_{i'=1}^n \lambda_i \lambda_{i'} Z_i Z_{i'} - \frac{2}{p} \sum_{i=1}^n \sum_{j=1}^p \lambda_i Z_i Z_j + \frac{1}{p^2} \sum_{j=1}^p \sum_{j'=1}^p Z_j Z_{j'}$$

$$E(\varepsilon(Z_v^*)^2) = \sum_{i=1}^n \sum_{i'=1}^n \lambda_i \lambda_{i'} Cov(Z_i, Z_{i'}) - \frac{2}{p} \sum_{i=1}^n \sum_{j=1}^p \lambda_i Cov(Z_i Z_j) + \frac{1}{p^2} \sum_{j=1}^p \sum_{j'=1}^p Cov(Z_j Z_{j'})$$

Derivando em relação ao ponderador tal como demonstrado na seção anterior encontramos a seguinte equação para a variância de krigagem:

$$\sigma_{\text{krig}}^2 = 2 \sum_{i'=1}^n \lambda_{i'} Cov(Z_i, Z_{i'}) - \frac{2}{p} \sum_{j=1}^p Cov(Z_i Z_j) \quad | \forall i$$

Em que:

$$\overline{Cov}(Z_i Z_j) = \frac{1}{p} \sum_{j=1}^p Cov(Z_i Z_j)$$

Logo temos:

$$\sigma_{krig}^2 = 2 \sum_{i=1}^n \lambda_i Cov(Z_i, Z_{i'}) - 2\overline{Cov}(Z_i Z_j) | \forall i$$

■

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 (8.2) demonstra como deve-se proceder para calcular a covariância média para cada amostra na krigagem de blocos.

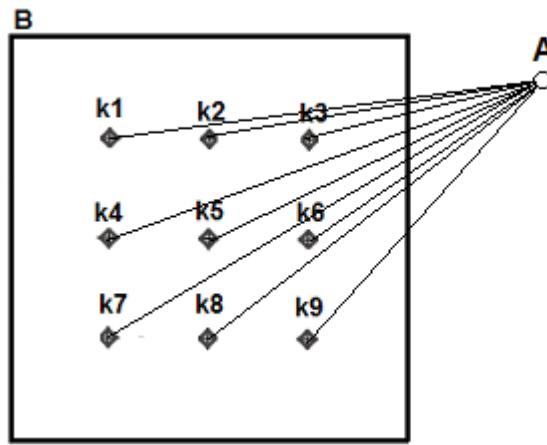


Figura 8.2: Covariância média como a média de covariâncias entre cada ponto estimado  $k$  dentro do bloco e o valor de cada amostra A

Logo o sistema de krigagem para blocos é demonstrado em (8.20)

$$\begin{pmatrix} Cov(Z_1, Z_1) & Cov(Z_1, Z_2) & \dots & Cov(Z_1, Z_n) & 1 \\ Cov(Z_2, Z_1) & Cov(Z_2, Z_2) & \dots & Cov(Z_2, Z_n) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(Z_n, Z_1) & Cov(Z_n, Z_2) & \dots & Cov(Z_n, Z_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ 1/2\mu \end{pmatrix} = \begin{pmatrix} \overline{Cov}(Z_0, Z_1) \\ \overline{Cov}(Z_0, Z_2) \\ \dots \\ \overline{Cov}(Z_0, Z_n) \\ 1 \end{pmatrix} \quad (8.20)$$

## 8.5 Influência nos pesos da krigagem

A krigagem é na verdade um estimador que não leva em consideração o valor da amostra, mas apenas sua correlação espacial e disposição no espaço das amostras. 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.

### 8.5.1 Influência do modelo de continuidade espacial nos pesos

A figura (8.3) 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 considerado um modelo de pepita puro, os pesos de krigagem são sempre os mesmos independente do posicionamento das amostras.

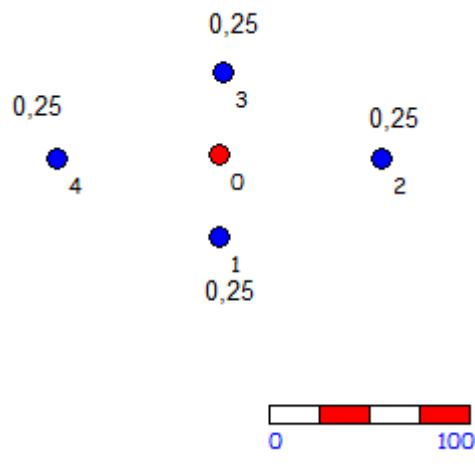


Figura 8.3: Demonstração dos pesos de krigagem para o posicionamento de amostras para um modelo de pepita puro

A figura (8.4) 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 (8.5) 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.

### 8.5.2 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 (8.6) demonstra os pesos de krigagem para um modelo esférico de variograma com um alcance de 63m

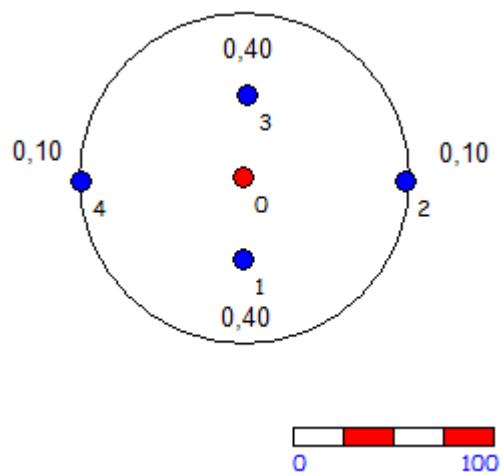


Figura 8.4: Demonstração dos pesos de krigagem para o posicionamento de amostras em um modelo esférico com alcance igual a  $2L$ . A linha cheia representa o alcance do variograma

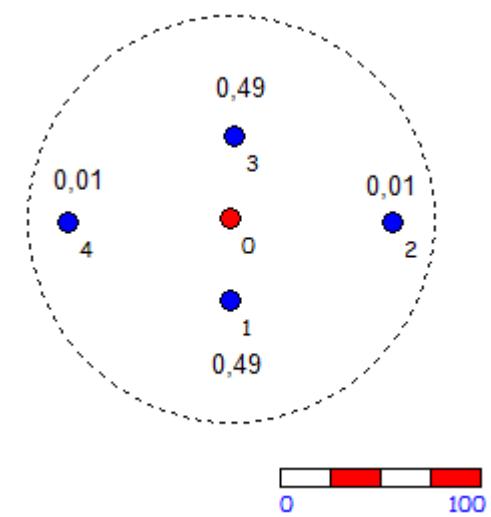


Figura 8.5: Demonstração dos pesos de krigagem para o posicionamento de amostras em um modelo gaussiano com alcance igual a  $1.5L$ . A linha hachurada demonstra o alcance prático do modelo de continuidade espacial.

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.

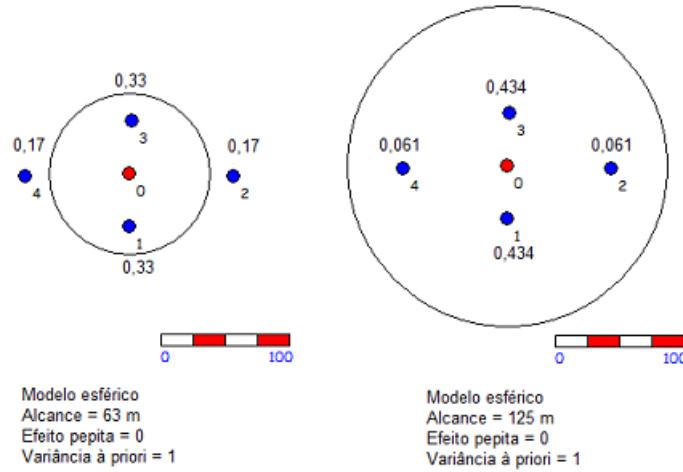


Figura 8.6: Influência dos parâmetros do variograma na krigagem. Efeito do alcance. Maiores alcances atribuem mais peso à amostras mais próximas

Quanto a variância à priori do variograma notamos segundo a figura (8.7) 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.

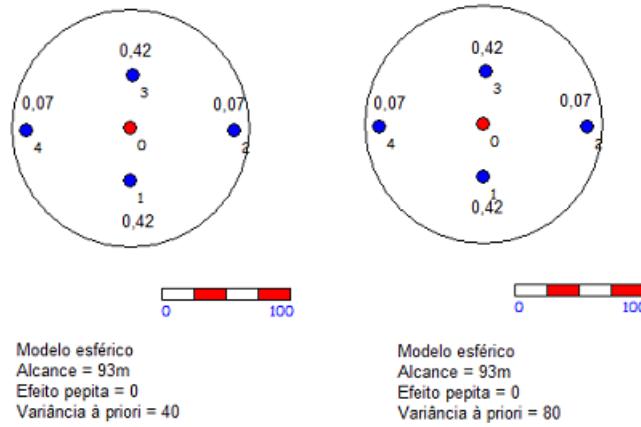


Figura 8.7: Influência dos parâmetros do variograma na krigagem. Efeito da variância a priori. Maiores valores de variância não alteram os pesos das amostras

Em último caso notamos segundo a figura (8.8) 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.

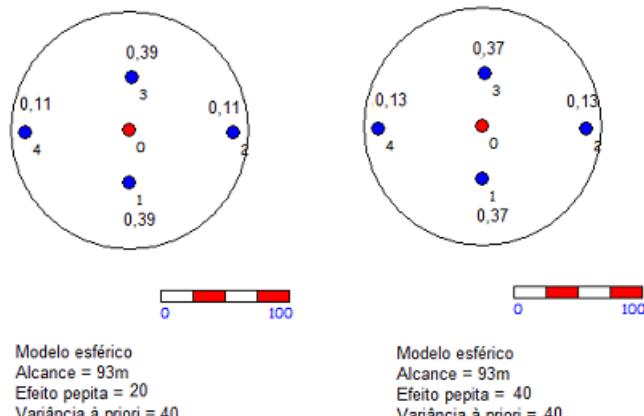


Figura 8.8: Influência dos parâmetros do variograma na krigagem. Efeito do alcance. Maiores alcances atribuem mais peso à amostras mais próximas

### 8.5.3 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 (8.9) demonstra esta relação.

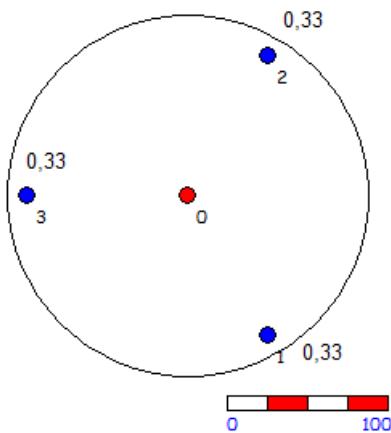


Figura 8.9: Influência no peso para distâncias iguais das amostras ao ponto estimado. Pesos iguais para um modelo de continuidade espacial isotrópico.

Para amostras agrupadas a tendência é produzir pesos iguais. A figura (8.10) demonstra esta situação.

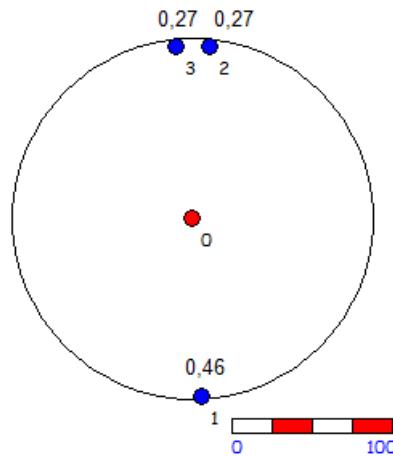


Figura 8.10: Influência de amostras agrupadas nos pesos da krigagem. Tendência de valores iguais para os pesos.

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 (8.11) demonstra o feito de blindagem das amostras e associação com pesos negativos.

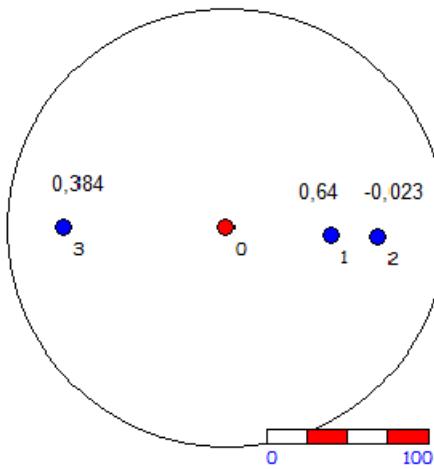


Figura 8.11: Influência da blindagem de amostras. Pesos negativos para amostras que estão encobertas por outras.

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-las tal como acontece com o modelo de efeito pepita puro. A figura (8.12) tende a demonstrar esta situação.

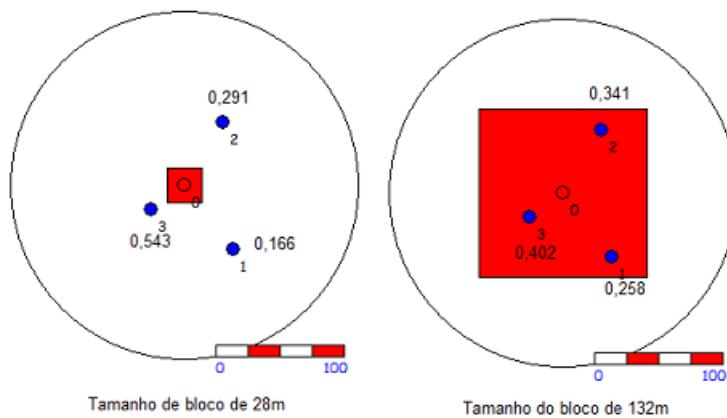


Figura 8.12: Influência do efeito de suporte do valor estimado. Blocos maiores tendem a normalizar o peso das amostras.

## 8.6 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 (8.13) demonstra esta situação.

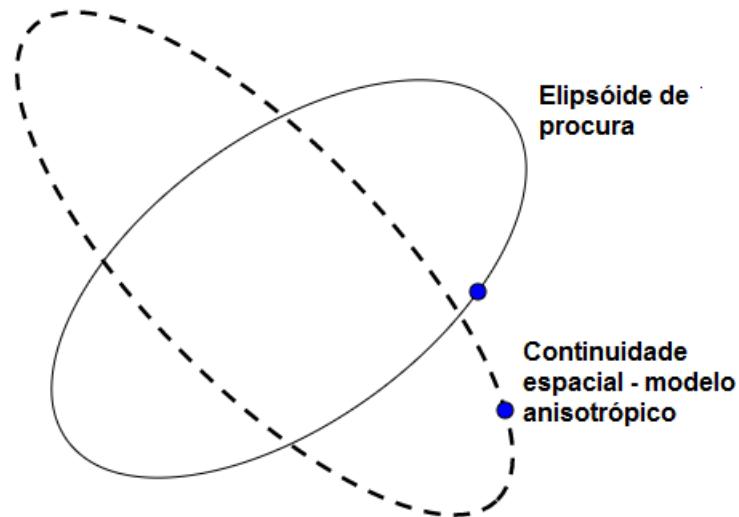


Figura 8.13: Estratégia de busca perpendicular à continuidade do fenômeno. Forma utilizada para garantir maior erraticidade aos dados estimados.

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 (8.14) 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.

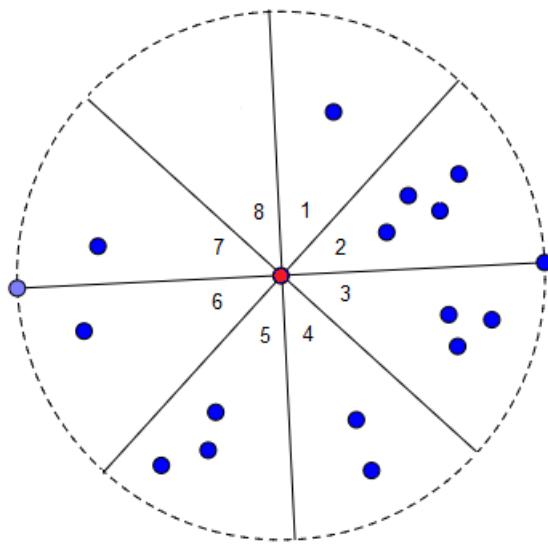


Figura 8.14: Estratégia de busca por octantes. Octantes 1,4,6 e 7 descartados por não possuírem o mínimo de amostras igual a 3

## 8.7 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. Comparaçāo da média global com a média das amostras
3. Análise de deriva de bandas do mapa
4. Validação Cruzada
5. Verificação de pesos negativos

### 8.7.1 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 fenômeno estudado, tal como continuidade espacial e regiões de maior ou menor valor da variável aleatória.

A figura (8.15) 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.

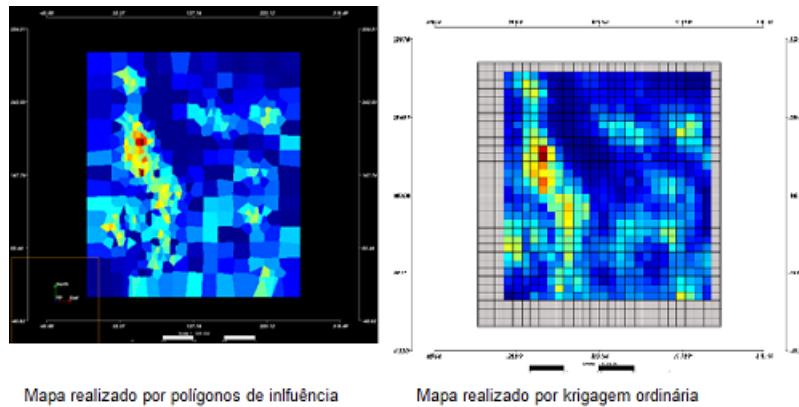


Figura 8.15: Comparação visual entre os valores da amostra por polígono de influência e dos blocos krigados.

### 8.7.2 Comparação da média global com a média das amostras

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.

### 8.7.3 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 (8.16). 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.

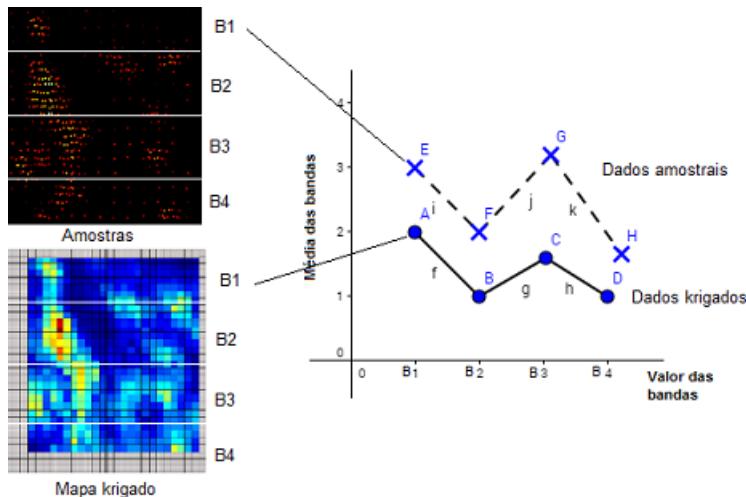


Figura 8.16: Exemplo da análise de deriva. Mama dos valores krigados e das amostras dividido em bandas e média tomada para cada banda em cada mapa demonstrado ao lado. Comportamento do gráfico semelhante para as duas situações. Descarte da hipótese de deriva.

#### 8.7.4 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.

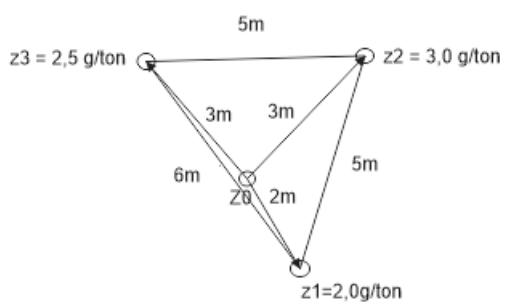
#### 8.7.5 Verificação de pesos negativos

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.

**Exercícios 8.1** A figura abaixo demonstra 3 pontos amostras e um ponto a ser estimado  $z_0$ . Considere o modelo de continuidade espacial como:

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

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







## 9. Mudança de suporte

### 9.1 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 transformar 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 (9.1), 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.

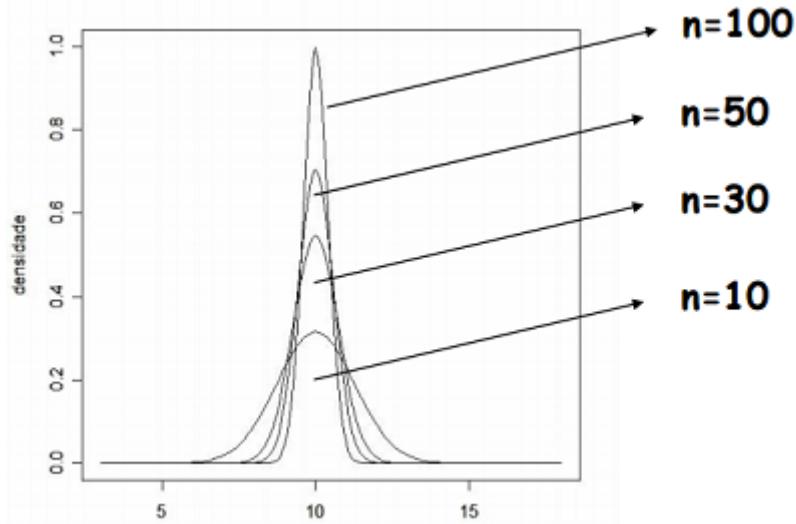


Figura 9.1: Figura demonstrando o efeito de suporte para um número crescente de amostras. O aumento do número de amostras tende a concentrar a função de densidade de probabilidade entorno do valor médio

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 (9.1)

$$f = \frac{\sigma_0^2 - \bar{\gamma}(V, V)}{\sigma_0^2} \quad (9.1)$$

Em que  $\bar{\gamma}(V, V)$  é o valor de variograma médio dentro do suporte V a ser corrigido e  $\sigma_0^2$  é o valor de variância à priori dos dados.

### 9.1.1 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 (9.2)

$$q' = \sqrt{f} * (q - m) + m \quad (9.2)$$

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.

### 9.1.2 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 resolução não linear para o problema, tal que cada quartil pode ser dado pela equação (9.3)

$$q' = aq^b \quad (9.3)$$

Em que  $a$  e  $b$  são constantes dadas por (9.4) e (9.5)

$$b = \sqrt{\frac{\ln(fCV^2 + 1)}{\ln(CV^2 + 1)}} \quad (9.4)$$

$$a = \frac{m}{\sqrt{fCV^2 + 1}} \left[ \frac{\sqrt{CV^2 + 1}}{m} \right]^2 \quad (9.5)$$

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.

## 9.2 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 (9.2) 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.

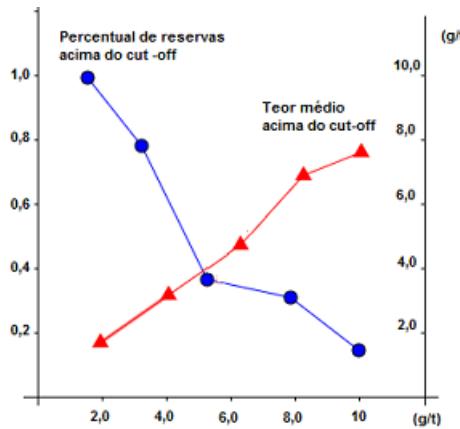


Figura 9.2: Curva teor e tonelagem para um depósito mineral genérico. Na curva azul temos o valor da proporção do depósito para um dado cut-off, enquanto na curva vermelha temos o teor médio acima de um cut-off.

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

### 9.2.1 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 (9.6)

$$P_{V \geq t_c} = 1 - P_{V=t_c} \quad (9.6)$$

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.

### **9.2.2 Curvas de teor e tonelagem a partir de distribuição de probabilidades contínuas 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.

### **9.2.3 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 (9.2.1).

### 9.2.4 Curvas de teor e tonelagem baseadas na estimativa dos blocos

Ao contrário do procedimento relatado em (9.2.3) 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 (9.3) demonstra a relação das curvas de cut-off para a proporção da jazida acima destes.

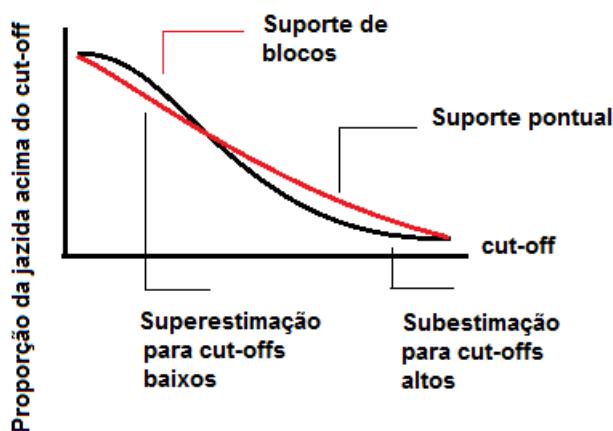


Figura 9.3: Demonstração da suavização da curva teor e tonelagem por mudança de suporte. Valores de cut-off mais baixos recebem maiores proporções, enquanto valores de cut-offs mais altos recebem maiores proporções

### 9.2.5 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.



## 10. Estimativa x Realidade

### 10.1 Introdução

A validação de qualquer estimativa somente pode ser feita comparando 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 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).

### 10.1.1 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 perspectiva 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.

### 10.1.2 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 (10.1) demonstra a eficiência do planejamento a longo prazo:

$$F_1 = \frac{\text{Planejado a curto prazo}}{\text{Planejado a longo prazo}} \quad (10.1)$$

A equação (10.2) demonstra o fator de eficiência para o planejamento de curto

prazo

$$F_2 = \frac{\text{Modelo de controle dos teores}}{\text{Planejado a curto prazo}} \quad (10.2)$$

A equação (10.3) demonstra a eficiência da informação passada pela mina

$$F_3 = \frac{\text{Reportado pela mina}}{\text{Modelo de controle dos teores}} \quad (10.3)$$

A equação (10.4) demonstra a eficiência da informação passada pela usina

$$F_4 = \frac{\text{Recebido pela usina}}{\text{Reportado pela mina}} \quad (10.4)$$

Esses fatores levam ao cálculo de alguns indicadores de performance tais como a precisão do planejamento de longo-prazo (long-term model) (10.5)

$$F_{LTM} = F_1 F_2 F_3 F_4 \frac{\text{Recebido pela usina}}{\text{Planejado a longo prazo}} \quad (10.5)$$

Ou o indicador do planejamento de curto prazo (short-term model) (10.6)

$$F_{STM} = F_2 F_3 F_4 \frac{\text{Recebido pela usina}}{\text{Planejado a curto prazo}} \quad (10.6)$$

A utilização de fatores de performance 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.

### 10.1.3 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:

$$G_a = C_r C_s G_e \quad (10.7)$$

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 (10.8)

$$CV_{G_a} \simeq \sqrt{\frac{s_{G_e}^2}{\bar{G}_e^2} + \frac{CV_{G_e}^2}{n} + CV_{C_1}^2 + CV_{C_2}^2} \quad (10.8)$$

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

#### 10.1.4 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 permaneçam 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 infor-

mações 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.

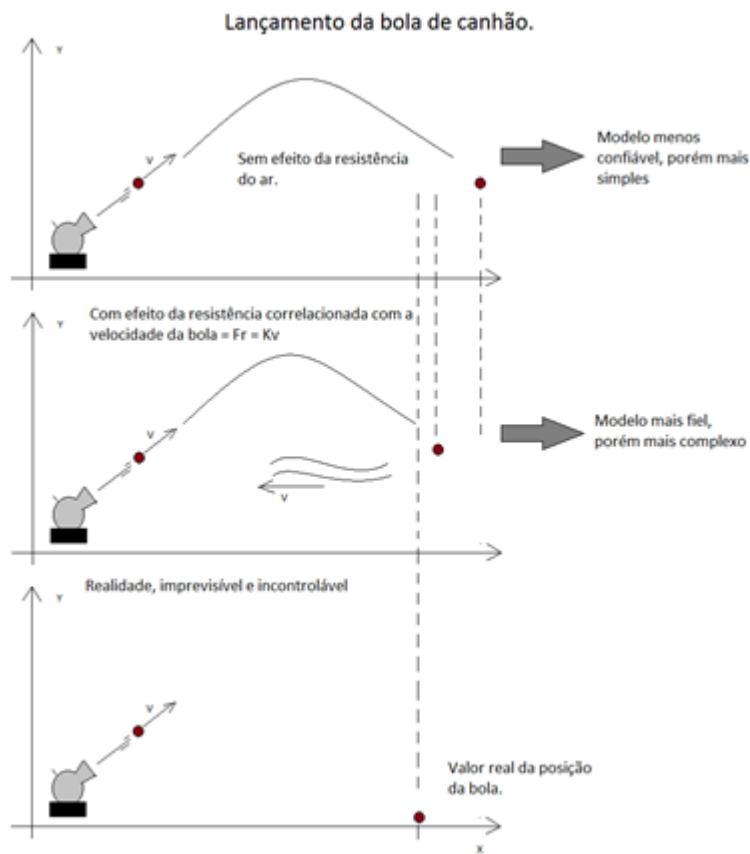


Figura A.1: Modelos físicos diferentes produzindo diferentes resultados. Incerteza sempre dependente do modelo escolhido

## 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:

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

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):

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

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  $f y$  sendo os expoentes das variáveis, ou seja  $a_1 y_1 + a_2 y_2^2 + \dots + a_n y_n^n$ .

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

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

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

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

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:

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

Temos então que:

$$f y(x_0) = \sum_{i=0}^n \lambda_i f y(x_i) \quad (\text{A.6})$$

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):

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

### 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:

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

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

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

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 covariâncias cruzadas.

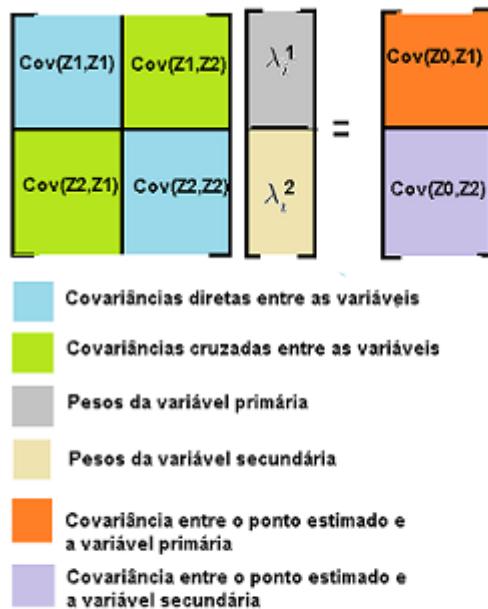


Figura A.2: Demonstração da matriz de cokrigagem para duas variáveis. Diferentes cores identificam as componentes da matriz

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 correacionalização.

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 correacionalização 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 iniciantes 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

Muitas vezes também é necessário realizar tarefas repetitivas, e se torna necessário resumir um número grande de instruções. Neste caso 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 [Isaaks and Srivastava \[1989\]](#). 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 Environment) 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 Environment) 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.

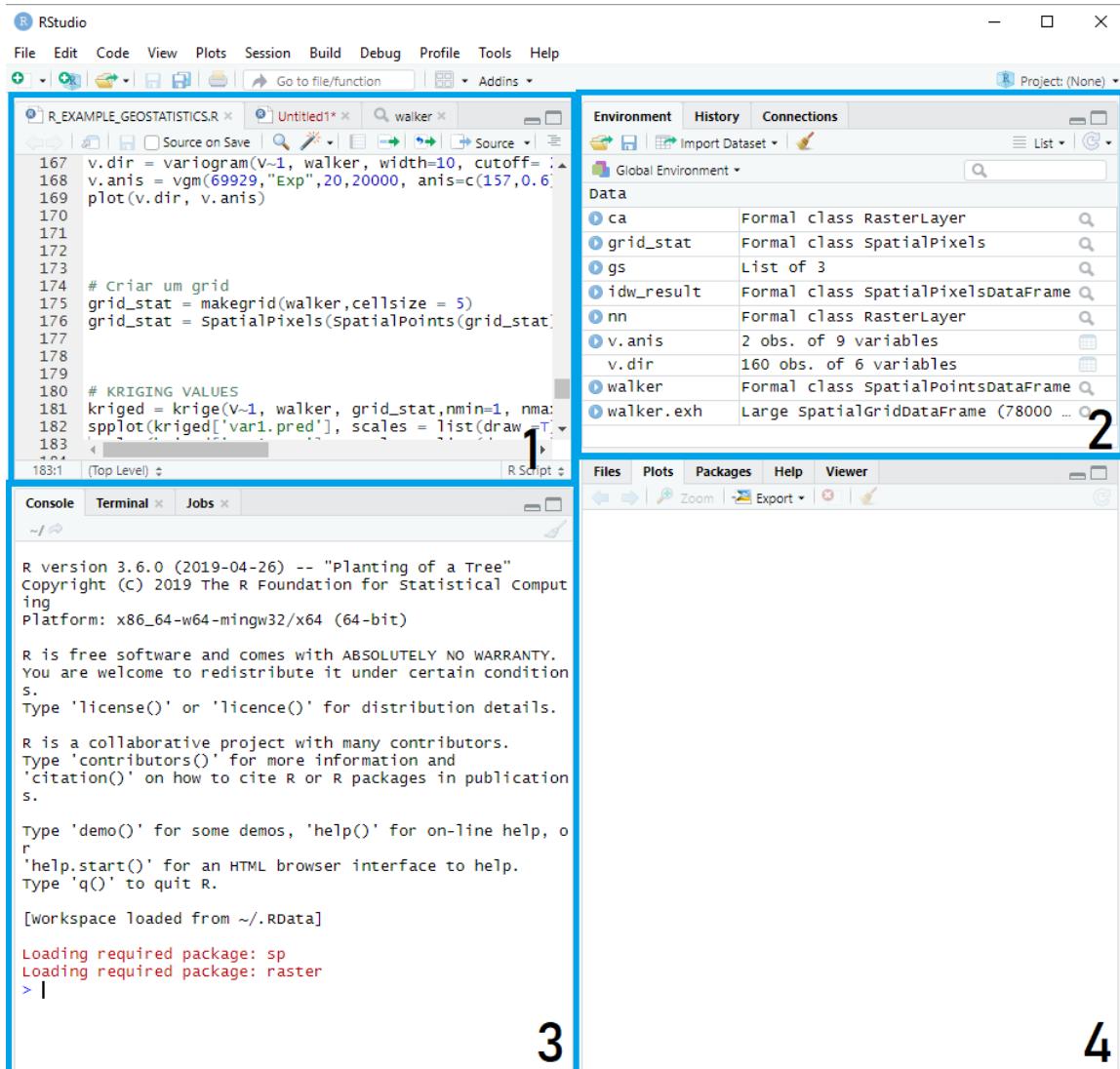


Figura B.1: Demonstração da janela do RStudio. 1 - Editor, 2-Variáveis de ambiente, 3-Console, 4- Output

## 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.

```

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 precedência 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
8 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 precedência de cada um destes é atribuído segundo a ordem estabelecida pela função. Abaixo encontramos algumas funções comuns do R.

```

1
2 log(3)      # logaritmo natural de 3

```

```

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

```

## 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 B.1. É importante lembrar que o operador `=` é associado à atribuição de uma variável ao contrário do operador `==` que representa igualdade.

Operador	Relação
<code>==</code>	Igualdade
<code>!=</code>	Diferente
<code>&gt;</code>	Maior
<code>&lt;</code>	Menor
<code>&gt;=</code>	Maior e igual
<code>&lt;=</code>	Menor e igual

Tabela B.1: Operadores relacionais no R

Em seguida são apresentados alguns resultados dos operadores relacionais para o R.

```

1
2  3 == 5          # Retorna FALSO
3  5 > 3          # Retorna VERDADEIRO
4  -2 < 7         # Retorna VERDADEIRO
5  abs(-5+3)== 2  # Retorna VERDADEIRO
6  4 != 4          # Retorna FALSO

```

## 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 Falso) 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 [B.2](#) 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 [B.3](#) 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

Em seguida demonstramos o código fonte de alguns operadores relacionais.

```

1
2 (5>4) & (3<7)      # Retorna VERDADEIRO
3 (5>4) & (3<2)      # Retorna FALSO
4 (7==4) || (2>3)     # Retorna FALSO
5 (3>2) || (2>3)     # Retorna VERDADEIRO

```

## 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. O código fonte a seguir demonstra como pedir ajuda para o R.

```

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

```

## 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`. O argumento `dependecies` permite com que o pacote instale qualquer outro tipo de dependência necessária para seu funcionamento, 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.1: 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")

```

Algumas operações podem ser realizadas com estes vetores. O código fonte a seguir demonstra algumas operações com 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

```

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. O código fonte a seguir demonstra como acessar valores de um vetor.

```

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

```

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.2: 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  If  (C > 5){  
2      print(C)  
3  }  
4  else{  
5      print(5)  
6  }
```

## 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 }
```

## 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 do código fonte a seguir.

```

1  y = sqrt(abs(-16)) # Retorna 4
2
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.3: 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 B.4 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 # Comandos para instalacao dos pacotes
2 install.packages(sp, dependences=T)
3 install.packages(gstat, dependences =T)
4 install.packages(geoR, dependences = T)
5
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
16
17 Lake

```

Listing B.4: 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 colunas no banco de dados. No caso de arquivos csv, o separador é a vírgula.

```

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

```

## 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))

```

O resultado do gráfico pode ser demonstrado na figura B.2. 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.

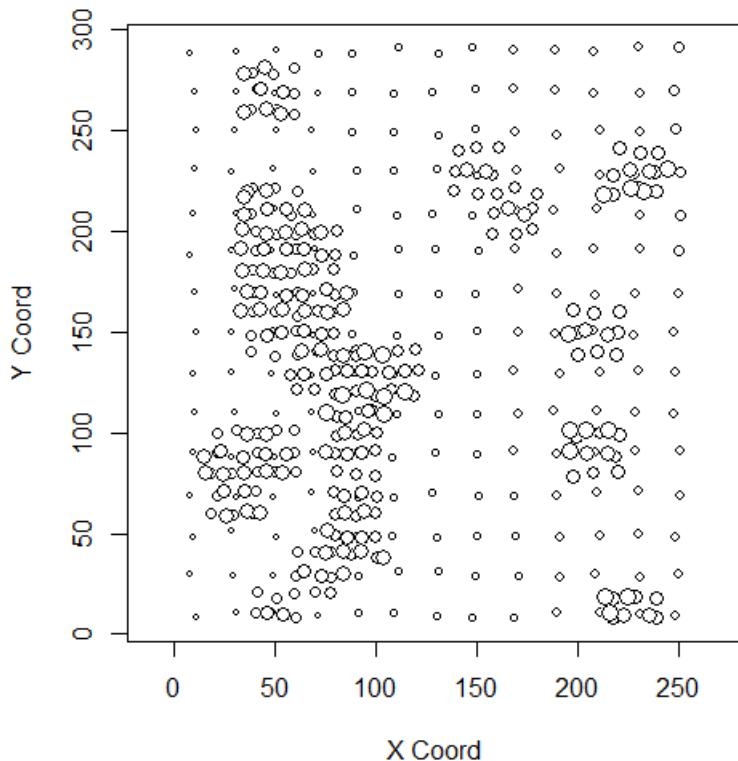


Figura B.2: Mapa de localização das amostras do Walker Lake

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 # SS PLOT DA VARIAVEL V

```

```

7  spplot(walker, c("V"), scales = list(draw =T))
8
9  # SS PLOT DA VARIAVEL V E U
10 spplot(walker, c("V", "U"), scales = list(draw =T))

```

A figura B.3 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.

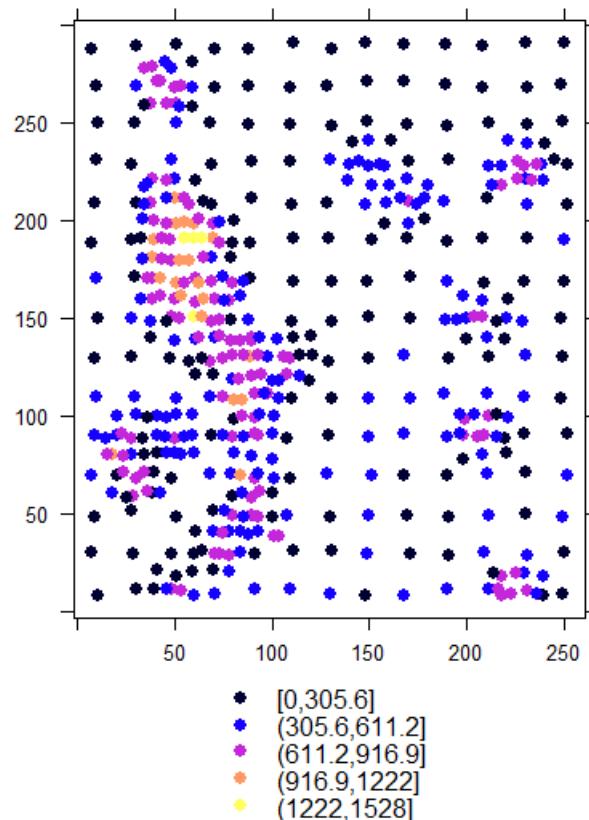


Figura B.3: Mapa de intervalos das amostras do Walker Lake, variável V

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

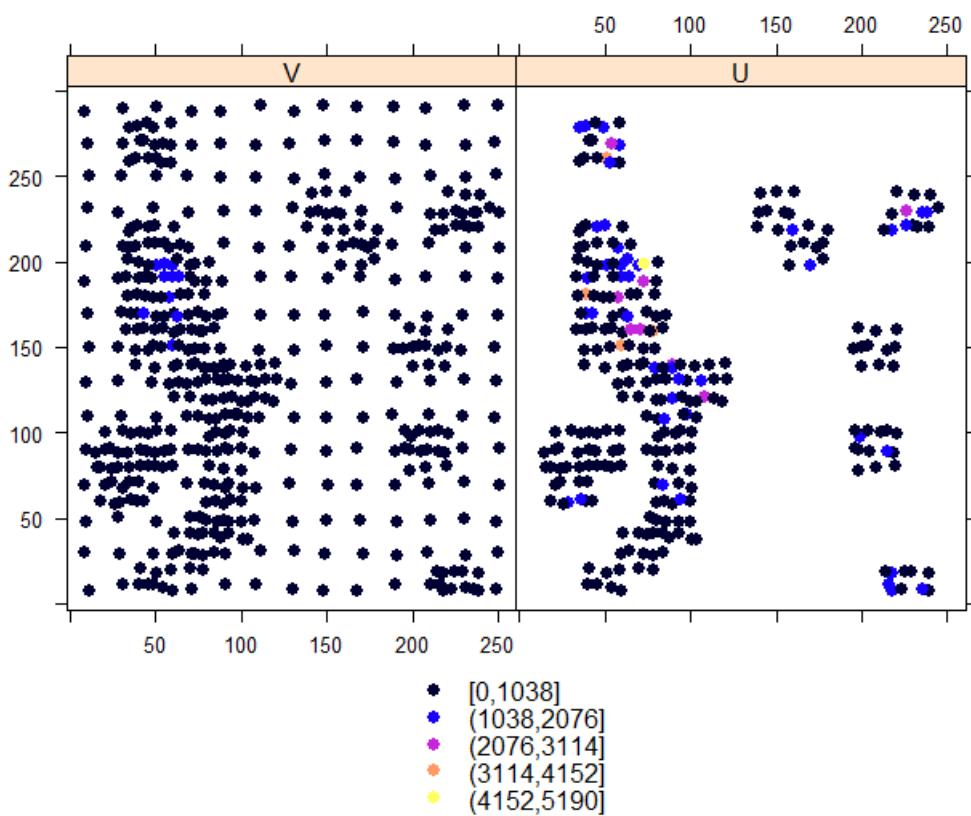


Figura B.4: Mapa de intervalos das amostras do Walker Lake, variável U e V

## 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. O código fonte a seguir demonstra os histogramas da variável U e V.

```

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=

```

```

8   U" , ylab= "Frequencia")
9
10 # Adicionar grafico da variavel V
11 hist(walker$V, main= "histograma da variavel V", breaks = 15, xlab=
12   "V" , ylab = "Frequencia")

```

A figura B.5 demonstra os histogramas gerados pelo código fonte. Notamos uma alta assimetria na variável U, enquanto a variável V demonstra valores mais espaçados entre si.

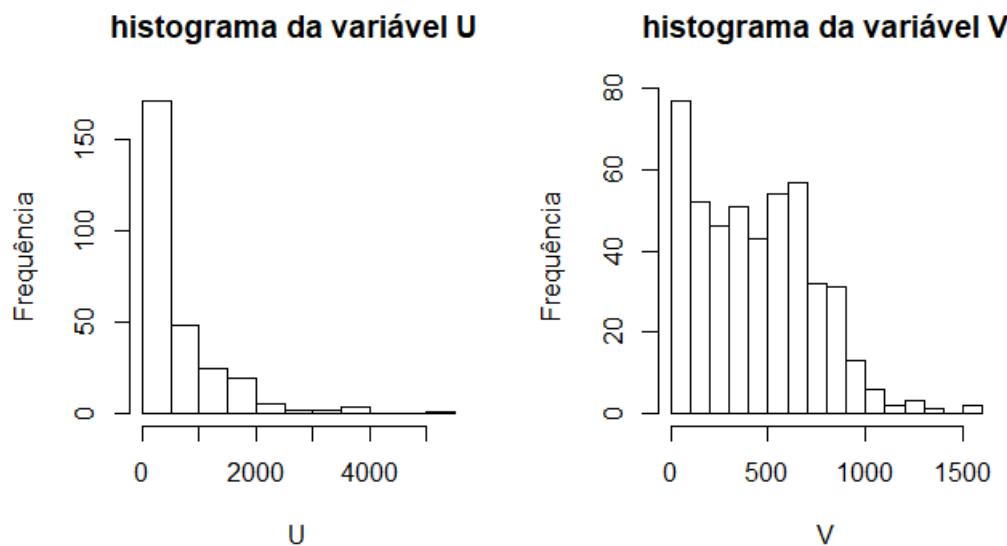


Figura B.5: Histogramas das variáveis U e V do Walker Lake

## 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. O código fonte a seguir demonstra a criação de gráficos de caixa para a variável U e V.

```

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")

```

A figura B.6 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.

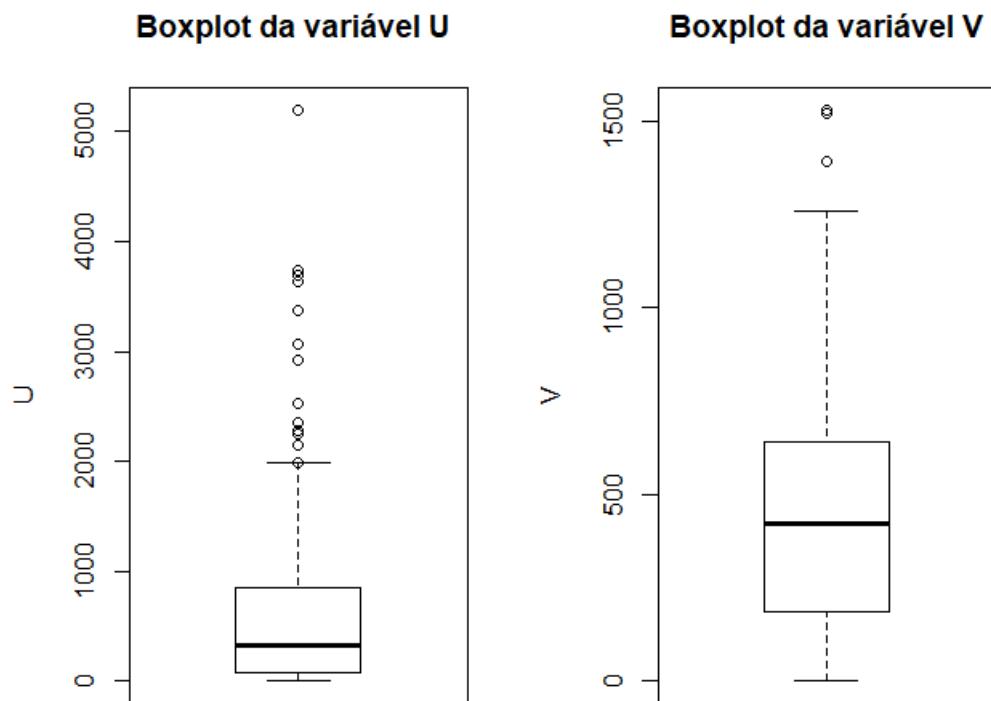


Figura B.6: Boxplot das variáveis U e V do Walker Lake

## 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 #Regressao Linear
2 linear = lm(walker$U ~ walker$V)
3 summary(linear)
4
5 # Plotagem do resultado
6 plot(walker$V, walker$U, xlab="V", ylab="U")
7 abline(linear)
```

Listing B.5: Criação de um vetor em R

O gráfico B.7 representa o modelo de regressão para as variáveis V e U. Um dos pontos acima do valor de 5000 parece demonstrar um valor outlier.

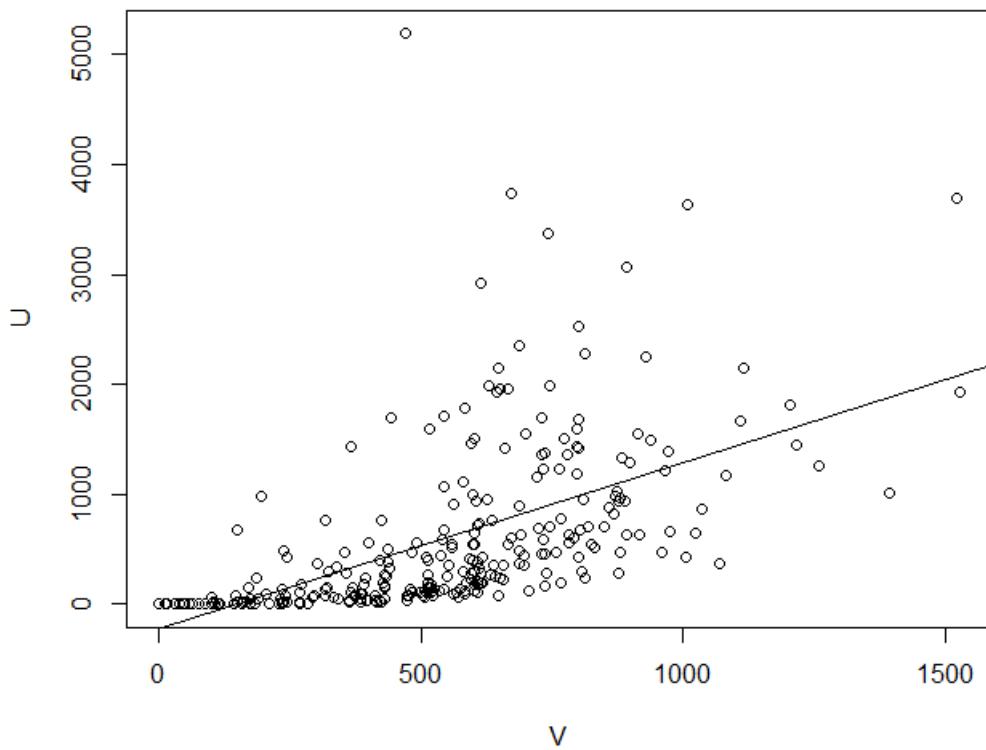


Figura B.7: Regressão linear das variáveis U e V do Walker Lake

O resultado da regressão pode ser expresso a partir da função `summary()`. Neste caso a regressão não apresenta uma boa performance, apresentando R-quadrado igual a 0.3016.

Call:

```
lm(formula = walker$U ~ walker$V)
```

Residuals:

Min	1Q	Median	3Q	Max
-1014.2	-386.8	-141.2	157.5	4704.0

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-225.2394	85.2053	-2.643	0.00868 **
walker\$V	1.5113	0.1384	10.923	< 2e-16 ***
---				

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 641.3 on 273 degrees of freedom
(195 observations deleted due to missingness)
Multiple R-squared: 0.3041, Adjusted R-squared: 0.3016
F-statistic: 119.3 on 1 and 273 DF, p-value: < 2.2e-16

```

## 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 # Criar um grid
3 grid_stat = makegrid(walker, cellsize = 5)
4 grid_stat = SpatialPixels(SpatialPoints(grid_stat))
5
6 # Calcular vizinho mais proximo
7
8 ca = raster(walker, res= 1)
9 gs = gstat(NULL, "V", V~1, walker, nmax=1)
10 nn = interpolate(ca, gs)
11 plot(nn, axes=T)

```

O resultado da interpolação pode ser compartilhado na figura B.8

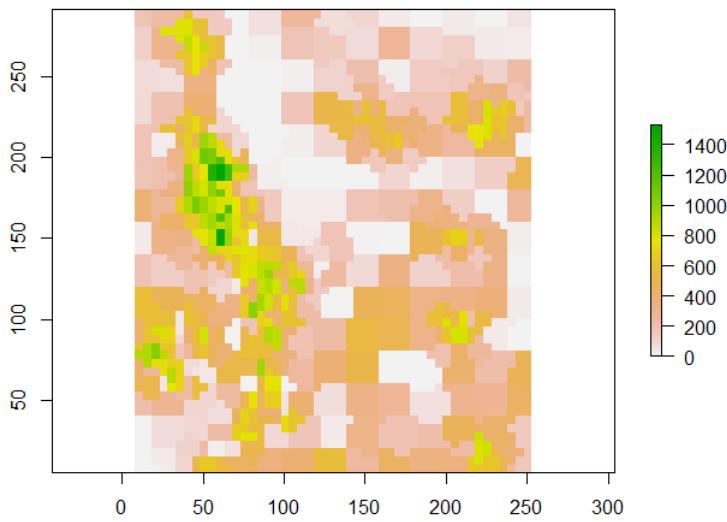


Figura B.8: Vizinho mais próximo Walker Lake - variável V

Para encontrarmos as estatísticas do vizinho mais próximo vamos utilizar o comando `summary()` e o comando `cellStats()` para obter os valores da média desagrupada e dos resultados do vizinho mais próximo.

```

1 summary(nn$V.pred)
2 cellStats(nn$V.pred, stat='mean')
3

```

Os resultados obtidos pelo R para o valor do vizinho mais próximo estão expressos a seguir. O valor da média desagrupada dos dados é em torno de 288, e ao final da interpolação vamos comparar as estatísticas finais com este valor.

```

> summary(nn$V.pred)
      V.pred
Min.       0.0
1st Qu.    79.4
Median     237.6
3rd Qu.    445.8
Max.     1528.1
NA's      0.0
> cellStats(nn$V.pred, stat='mean')
[1] 287.9967

```

## 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)

```

A figura B.9 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.

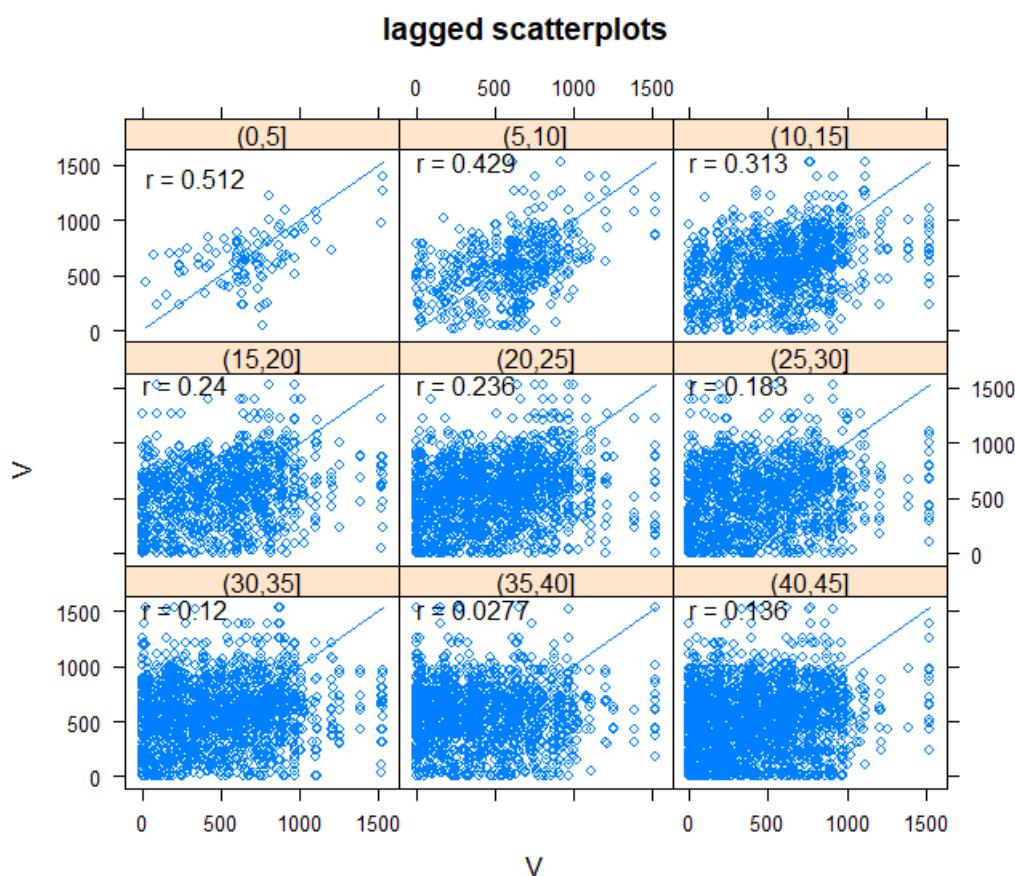


Figura B.9: Hscatterplot para a variável V

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 utilizamos 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 #Variancia da variavel V - 89929.4
2 var(walker$V)
3
4 #Variogramas experimentais
5
6 v.dir = variogram(V~1, walker, width=10, cutoff= 200, tol.hor=45,
alpha = (0:7)*22.5 )

```

Listing B.6: Criação de um vetor em R

A figura B.10 demonstra os variogramas experimentais para a variável V do depósito Walker Lake. Notamos que a direção de azimute 157.5 graus é a direção de maior continuidade do depósito mineral. Podemos concluir isto não apenas inferindo um alcance para esta direção, mas também observando no mapa dos dados da figura B.8. Para mapear a direção de maior continuidade pode-se realizar variogramas em diferentes direções, verificando um por um, qual é aquele que possui maior alcance.

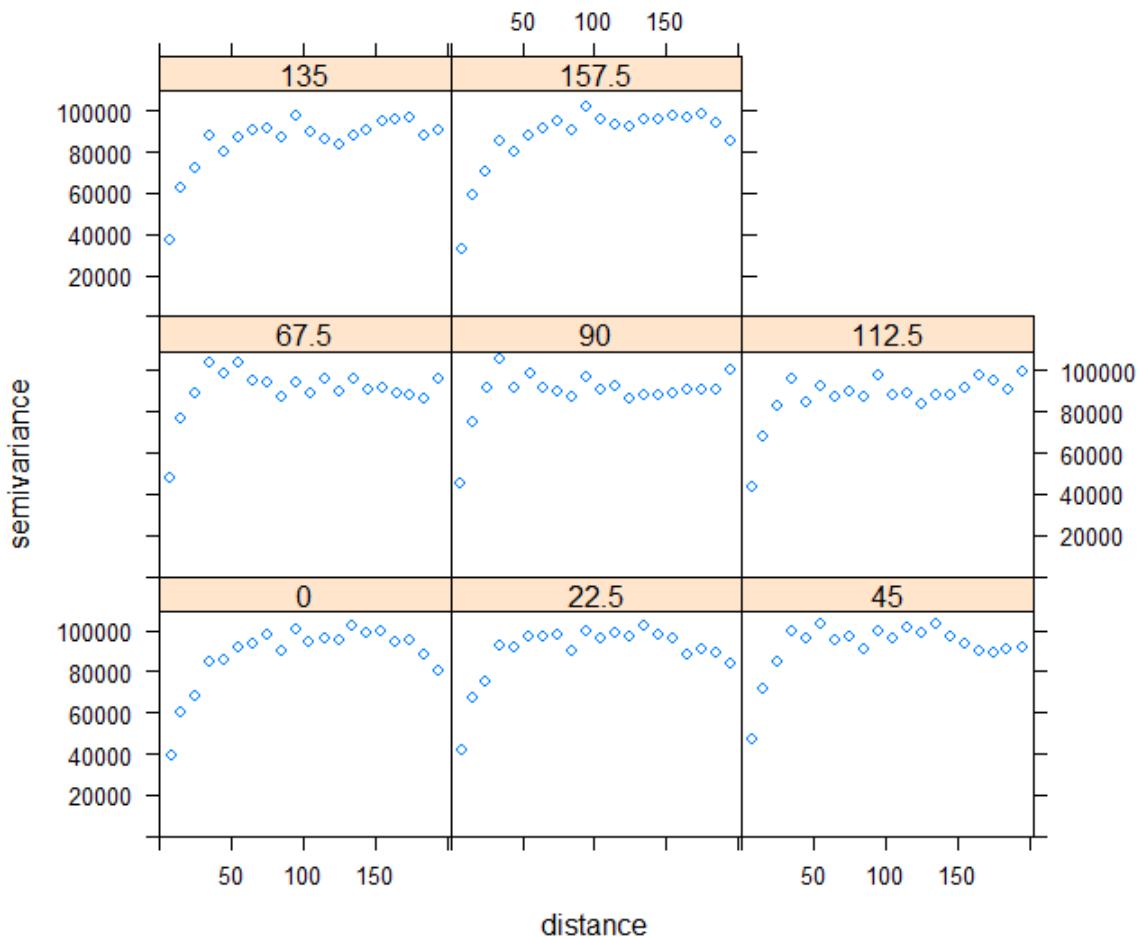


Figura B.10: Variogramas experimentais gerados

A modelagem de variogramas individuais pode ser realizada a partir da biblioteca geoR utilizando o comando eyefit. Este comando permite modelar apenas um variograma manualmente a partir de uma janela para o ajuste. Para realizar o cálculo criamos um objeto do variograma experimental utilizando a função variog() e atribuímos os argumentos relacionados com o variograma, tais como direção (em radianos), tolerância angular (em radianos), o argumento uvec, que representa o vetor com distâncias a serem calculados os valores do variograma e a máxima distância de cálculo. 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 #Criar variograma experimental utilizando a biblioteca geoR
3 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)

```

```

4
5 #Fitar manualmente o variograma
6 ve.eye = eyefit(var)
7
8 #Transformar o modelo fitado em um modelo do gstat
9 ve.fit = as.vgm.variomodel(v.eye[[1]])

```

A figura B.11 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.

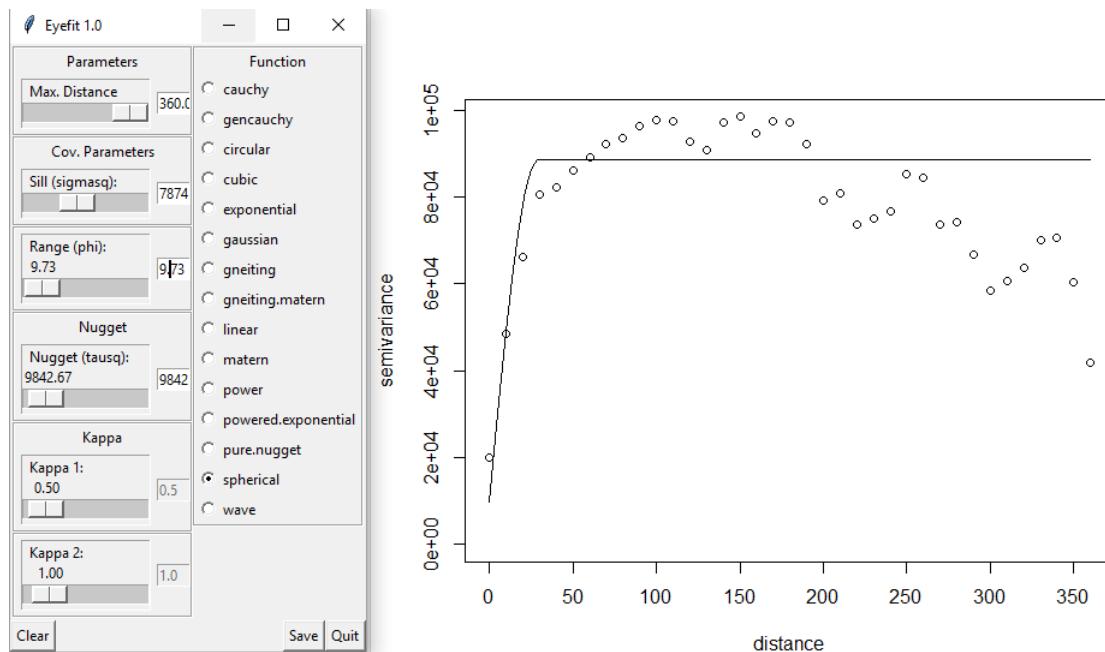


Figura B.11: Ajuste do modelo de variograma utilizando a função eyefit

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 B.5 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  #Variância da variável V
4  var(walker$V)
5

```

```
6 #Variogramas experimentais
7
8 v.dir = variogram(V~1, walker, width=10, cutoff= 200, tol.hor=45,
9   alpha = (0:7)*22.5 )
10
11 # Modelagem da primeira estrutura
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)
```

A figura B.12 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.

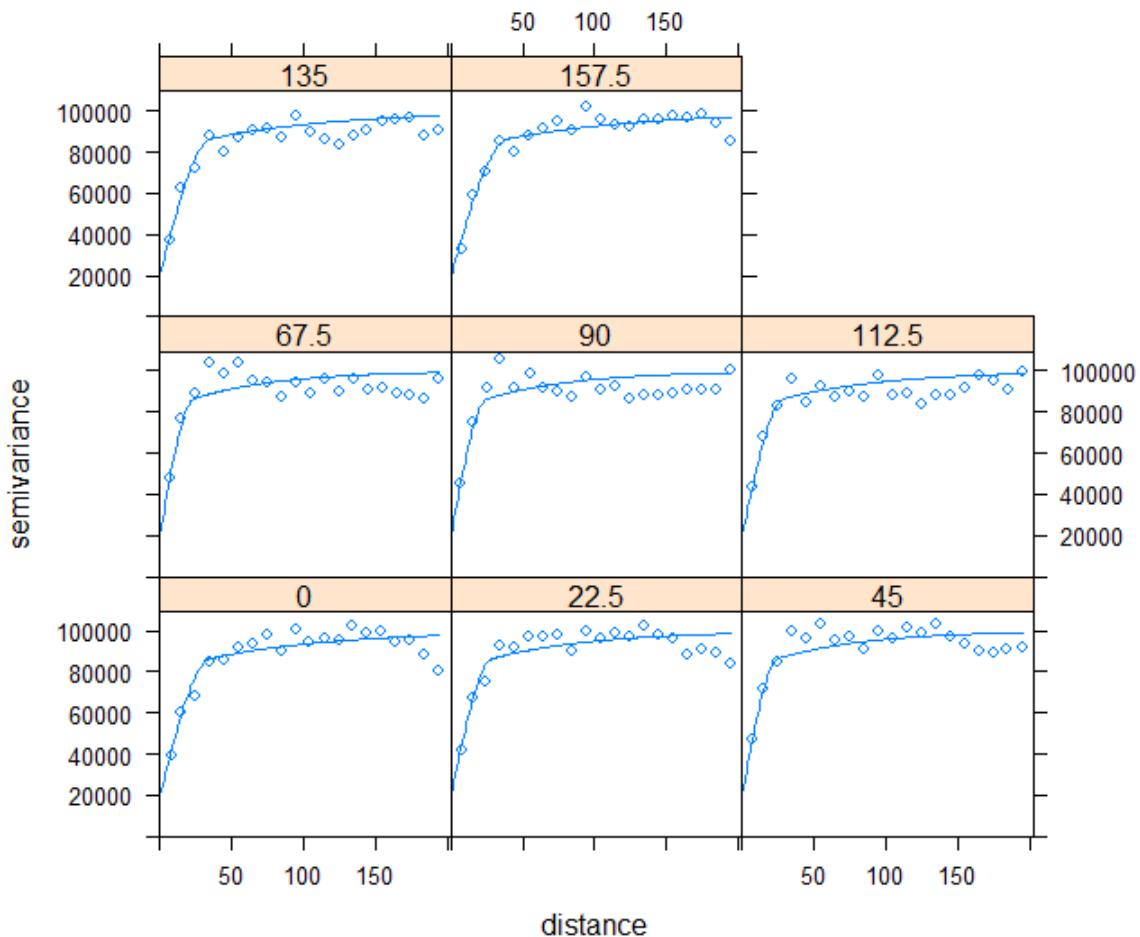


Figura B.12: Ajuste do modelo para diferentes direções utilizando a função vgm

## 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 # Validacao cruzada
2 cv = krige.cv(V~1, walker, v.anis2, nmin= 3, nmax=10, maxdist=100,
   nfold=20)
3
4 # Sumario estatistico da validacao
5 summary(cv)
6
7 #Plotagem dos resíduos da validacao cruzada
8 bubble(cv[ "residual"])
9

```

A figura B.13 demonstra os erros residuais a partir da validação cruzada do depósito Walker Lake.

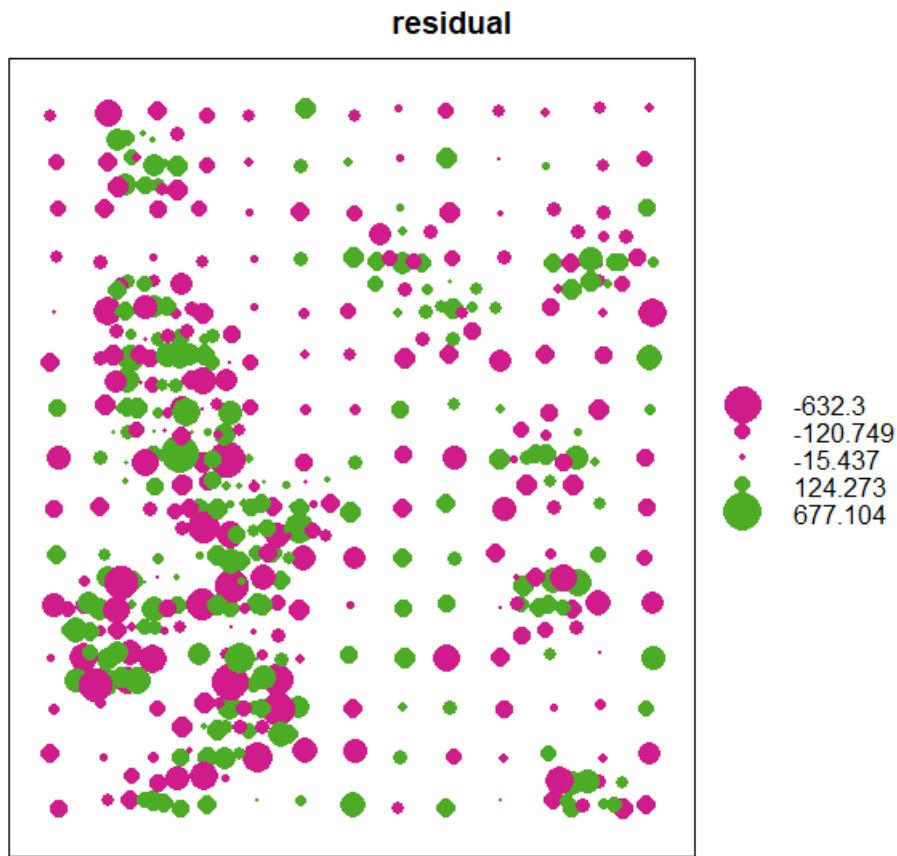


Figura B.13: Erros da validação cruzada demonstrados em um gráfico de bolhas

O resumo da validação cruzada, utilizando o comando `summary()`, para a variável V, pode ser verificado a seguir. 5 colunas são demonstradas com o resumo dos dados,

incluindo os valores dos resíduos e dos valores preditos da variável.

```
> summary(cv)
Object of class SpatialPointsDataFrame
Coordinates:
    min   max
X     8 251
Y     8 291
Is projected: NA
proj4string : [NA]
Number of points: 470
Data attributes:
      var1.pred      var1.var      observed      residual      zscore
Min.   : 25.9   Min.   :34698   Min.   : 0.0   Min.   :-644.93   Min.   :-2.98226
1st Qu.: 280.6  1st Qu.:42693   1st Qu.: 184.6  1st Qu.:-129.58  1st Qu.:-0.52658
Median : 427.1  Median :47905   Median : 424.0  Median : -14.62  Median : -0.05799
Mean   : 445.5  Mean   :55645   Mean   : 435.3  Mean   : -10.24  Mean   : -0.02241
3rd Qu.: 588.1  3rd Qu.:68170   3rd Qu.: 640.9  3rd Qu.: 125.54  3rd Qu.: 0.56901
Max.   :1163.5  Max.   :98031   Max.   :1528.1  Max.   : 676.99  Max.   : 3.30583
```

## 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 (maxdist) 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.
8   anis2)
9
9 # Plotar a variavel estimada
10 spplot(kriged[ 'var1.pred' ], scales = list(draw =T))
```

```
11  
12 # Plotar a variância de krigagem  
13 spplot(kriged[ 'var1.var' ] , scales = list(draw =T))  
14 summary(kriged)
```

O gráfico B.14 demonstra o valor krigado da variável V a partir da estratégia de busca e do variograma ajustado.

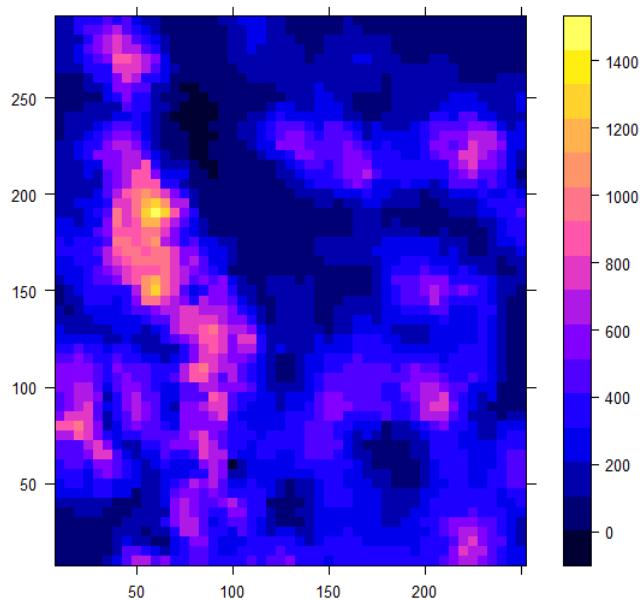


Figura B.14: Krigagem da variável V do depósito Walker Lake

O gráfico B.15 demonstra a variância de krigagem da variável V a partir da estratégia de busca e do variograma ajustado.

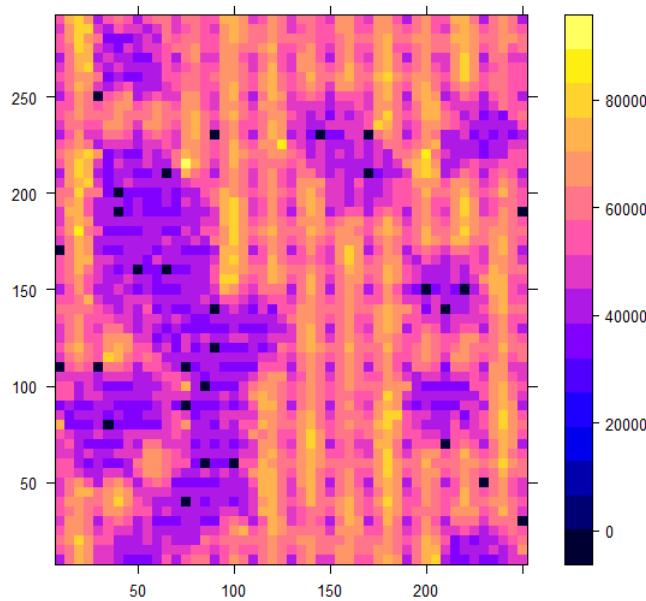


Figura B.15: Variância de krigagem da variável V do depósito Walker Lake

O resultado das estatísticas é demonstrado em seguida. Não houve previsões de valores negativos e o valor médio está próximo do vizinho mais próximo.

```
> summary(kriged)
Object of class SpatialPixelsDataFrame
Coordinates:
    min     max
x1 7.5 252.5
x2 7.5 292.5
Is projected: NA
proj4string : [NA]
Number of points: 2793
Grid attributes:
    cellcentre.offset  cellsize  cells.dim
x1                  10        5        49
x2                  10        5        57
Data attributes:
    var1.pred          var1.var
Min.   : 0.0   Min.   : 0
1st Qu.: 125.7 1st Qu.:44914
Median : 256.7  Median :55883
```

---

```
Mean      : 295.5    Mean      :54525
3rd Qu.: 410.7    3rd Qu.:63748
Max.     :1432.9    Max.     :89477
```



## C. Geoestatística utilizando o GSLib

### C.1 Introdução

O GSLIB (Geostatistical Software Library) é um pacote de programas desenvolvidos em linguagem Fortran junto à Universidade de Stanford, nos Estados Unidos da América, sob a direção do professor André G. Journel e do professor Calyton V. Deutsch. Atualmente os pacotes de geoestatística para o GSLIB90 estão disponíveis no link <http://www.gslib.com/>, na aba Download. Além dos executáveis, é possível também baixar os códigos fontes em Fortran para cada um dos programas.

Apesar de ser voltado principalmente para alunos de pós-graduação e pesquisadores, por ser um software gratuito muitas empresas ainda utilizam suas rotinas para análise geoestatística. O GSLIB pode ser considerado um dos melhores programas à disposição para análise espacial, permitindo o uso de algoritmos em diversas áreas como variografia, krigagem e simulação. Os resultados gráficos gerados pelo GSLIB são convertidos em arquivos PostScript. Para a visualização e impressão dos dados podemos utilizar o programa Ghostscript obtido no link <https://www.ghostscript.com/>. Outro programa importante para a visualização é o ghostviewer e pode ser obtido no seguinte link <http://www.ghostgum.com.au/software/gsview.htm>.

O GSLIB geralmente não se apresenta de forma amigável para o usuário pois são baseados em antigas plataformas como o DOS para windows, e não possuem uma interface gráfica.

## C.2 A execução do GSLIB

Para rodar o GSLib é necessário possuir os arquivos executáveis(.exe) em uma pasta no computador. Quando executado, o arquivo requisitará a criação de um arquivo (.par) na mesma pasta. Enquanto o arquivo executável é responsável pela execução do programa, o arquivo par é responsável por inserir os parâmetros necessários para o algoritmo.

## C.3 Entrada de dados

Os dados de entrada do GSLIB são escritos em arquivos ASCII em um formato simplificado. Por convenção, os dados orientados para Leste são considerados no eixo X positivamente, enquanto os dados orientados para o Norte são considerados no eixo Y positivamente. Um banco de dados geoestatístico deve possuir a orientação da coordenada da amostra juntamente com o valor dos atributos associados. A seguir é demonstrado um exemplo de banco de dados do Walker Lake. Primeiramente informamos o nome do banco de dados, sem seguida o número de variáveis totais no banco, então fornecemos o nome de cada variável em cada linha e por último os dados são fornecidos em cada linha, separados com espaços. Para o funcionamento correto do GSLIB é necessário que a separação de decimais dos valores seja realizada por ponto. Os programas do GSLIB permitem a filtragem de valores faltantes. Isso pode ser feito dentro do arquivo par pelo método trimming limits. Como a filtragem é realizada nos extremos dos dados geralmente se atribui um valor negativo muito grande como -999, filtrando apenas os valores acima de 0. É importante também remover quaisquer linhas no final do banco de dados vazia.

```
DataSet
4
X
Y
Cu (%)
Au (PPM)
11 8 0 -999
8 30 0 -999
9 48 2.243999939 -999
8 69 4.343999939 -999
9 90 4.121000061 -999
10 110 5.872000122 -999
9 129 1.923000031 -999
11 150 0.312999992 -999
```

```

10 170 3.885 -999
8 188 1.746000061 -999
9 209 1.878000031 -999
10 231 0.820999985 -999
11 250 0.810999985 -999
10 269 1.243000031 -999
8 288 1.88 -999
31 11 0.287000008 -999

```

## C.4 Exemplos de aplicação do GSLIB

A seguir realizaremos alguns exemplos do uso do GSLIB no banco de dados do Walker Lake. O banco de dados apresenta coordenadas planas X e Y e mais duas variáveis de teor de cobre e ouro artificialmente criadas a partir de um banco de dados de topografia, de uma região de Nevada no Canadá.

### C.4.1 Criando um histograma com o HISTPLT

Para gerar o histograma utilizamos o programa HISTPLTt. O texto a seguir demonstra o arquivo de parâmetros.

```

Parameters for HISTPLT
*****
START OF PARAMETERS:
Walker_Lake.txt      -file with data
3                      -columns for variable and weight
-1.0      1.0e21        -trimming limits
histograma_Cu.ps     -file for PostScript output
0.0       20.0          -attribute minimum and maximum
-1.0                  -frequency maximum (<0 for automatic)
20                     -number of classes
0                      -0=arithmetic, 1=log scaling
0                      -0=frequency, 1=cumulative histogram
0                      -number of cum. quantiles (<0 for all)
2                      -number of decimal places (<0 for auto.)
Histograma Cu        -title
1.5                   -positioning of stats (L to R: -1 to 1)
0                      -reference value for box plot

```

Cada linha representa os seguintes parâmetros do arquivo de parâmetros.

1. Endereço do arquivo do banco de dados
2. Número da coluna da esquerda para direita da variável e dos pesos para a geração do histograma
3. Limites para o uso dos dados. Limite inferior (-1.0) e limite superior ( $10^{21}$ )
4. Nome do arquivo postscript de saída do histograma
5. Número mínimo e máximo demonstrado no histograma
6. Máxima frequência
7. Número de classes do histograma
8. Escala aritmética ou logarítmica
9. Histogramas de frequência ou acumulado
10. Número de quantis para o histograma acumulado
11. Número de decimais
12. Título do gráfico
13. posicionamento das estatísticas (-1 esquerda, 1 direita)
14. ponto de referência para plotagem do gráfico de caixa

As figuras [C.1](#) e [C.2](#) representam os histogramas das variáveis cobre e ouro. Podemos notar a assimetria característica das duas variáveis, principalmente na de ouro.

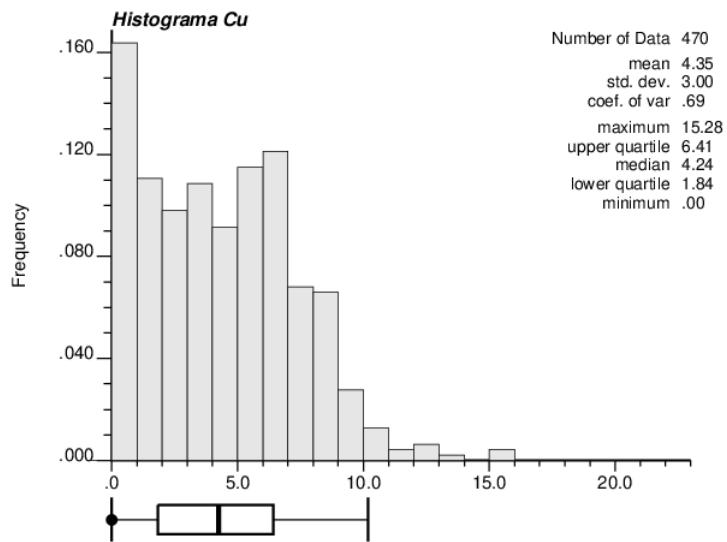


Figura C.1: Histograma do Cobre para o depósito do Walker Lake

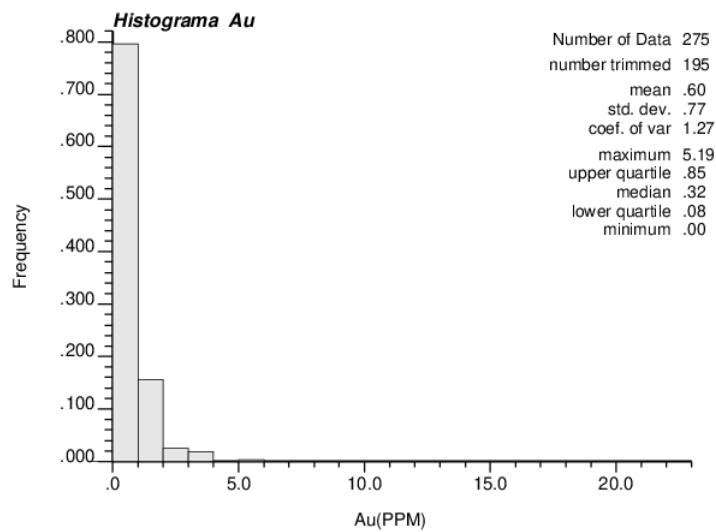


Figura C.2: Histograma do Ouro para o depósito do Walker Lake

### C.4.2 Criando um gráfico de dispersão com o SCATPLT

Gráficos de dispersão são uma boa forma de se avaliar a dependência entre duas variáveis X e Y diferentes. Para realizar o gráfico de dispersão é demonstrado a seguir o arquivo de parâmetros utilizado no depósito Walker Lake do programa SCATPLT.

```
Parameters for SCATPLT
*****
START OF PARAMETERS:
Walker_Lake.txt          -file with data
3   4   0   0              -columns for X, Y, wt, third var.
0   1.0e21                 -trimming limits
Scatter_walker.ps         -file for Postscript output
0.0    18    0             -X min and max, (0=arith, 1=log)
0.0    7     0             -Y min and max, (0=arith, 1=log)
1                  -plot every nth data point
0.5                 -bullet size: 0.1(sml)-1(reg)-10(big)
0.0    2.0                -limits for third variable gray scale
Scatterplot Cu(%) /Au(ppm) -title
```

Em seguida é demonstrado o significado de cada parâmetro no arquivo .par.

1. Endereço do arquivo do banco de dados
2. Número da coluna X, da coluna Y, os pesos de desagrupamento e uma terceira variável
3. Limites para o uso dos dados. Limite inferior (-1.0) e limite superior ( $10^{21}$ )
4. Nome do arquivo postscript de saída do histograma
5. Número mínimo, máximo da variável X e escolha da escala (0 = aritmética, 1 = logarítmica)
6. Número mínimo, máximo da variável Y e escolha da escala (0 = aritmética, 1 = logarítmica)
7. Plotar a cada n pontos
8. Tamanho do ponto no gráfico

## 9. Limites para a terceira variável

## 10. Título do gráfico

A figura C.3 demonstra o resultado para as variáveis cobre e ouro do depósito Walker Lake. Nota-se que o coeficiente de correlação de Pearson é baixo, apresentando valor de 0.551. Também é apresentado na figura o coeficiente de rank de 0.762, demonstrando que pode haver uma correlação não necessariamente linear entre as variáveis.

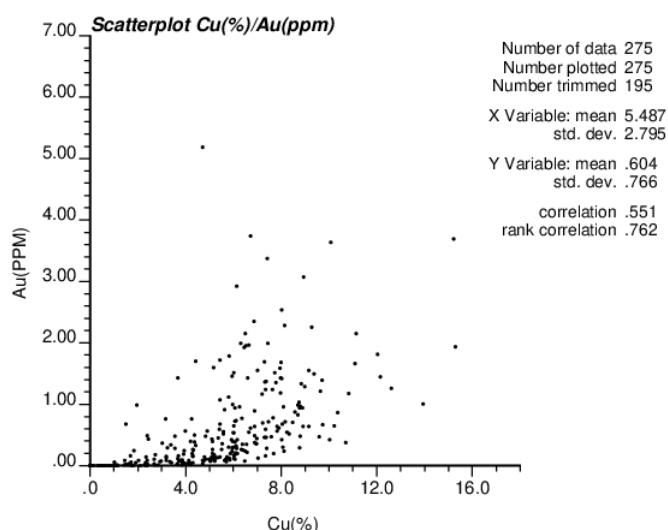


Figura C.3: Diagrama de dispersão do Cobre e Ouro para o depósito do Walker Lake

### C.4.3 Criando um mapa de localização com o LOCMAP

Os mapas de localização são uma forma visual interessante para se determinar os padrões de localização das amostras, tal como as regiões onde ocorrem anomalias positivas e negativas. A seguir é demonstrado o arquivo de parâmetros do arquivo .par do programa LOCMAP.

```
Parameters for LOCMAP
*****
START OF PARAMETERS:
Walker_Lake.txt          -file with data
1   2   3                  - columns for X, Y, variable
```

```

-1.0      1.0e21          - trimming limits
Mapa_Cu.ps                         -file for PostScript output
0.0       260               -xmn, xmx
0.0       300               -ymn, ymx
0                     -0=data values, 1=cross validation
1                     -0=arithmetic, 1=log scaling
1                     -0=gray scale, 1=color scale
0                     -0=no labels, 1=label each location
0.01     10.0    1.        -gray/color scale: min, max, increm
0.3                  -label size: 0.1(sml)-1(reg)-10(big)
Mapa Cu (%)           -Title

```

Em seguida são demonstrados os significados dos parâmetros em cada linha do arquivo .par.

1. Endereço do arquivo do banco de dados
2. Número da coluna X, da coluna Y, os pesos de desagrupamento e uma terceira variável
3. Limites para o uso dos dados. Limite inferior (-1.0) e limite superior ( $10^{21}$ )
4. Nome do arquivo postscript de saída do histograma
5. Número mínimo, máximo da variável X
6. Número mínimo, máximo da variável Y
7. Plotar os valores dos dados ou da validação cruzada (0=dados, 1=validação cruzada)
8. Plotar dados em escala aritmética ou logarítmica (0=aritmética, 1=logarítmica)
9. Escala de cor cinza ou colorida (0=cinza, 1= colorida)
10. adição de rótulos nos dados (0= sem rótulos, 1= rótulo dos dados)
11. valor mínimo, máximo e incremento da escala
12. tamanho da fonte dos rótulos
13. Título do gráfico

A figura C.4 demonstra o resultado gráfico do mapa de localização das amostras do Walker Lake. Podemos notar um corpo de maior extensão no flanco oeste e corpos menores ao leste. A orientação do corpo maior possui alongamento na direção sudeste. Os valores mais intensos de teor de cobre se encontram dentro do corpo geológico, enquanto as bordas se apresentam mais pobres.

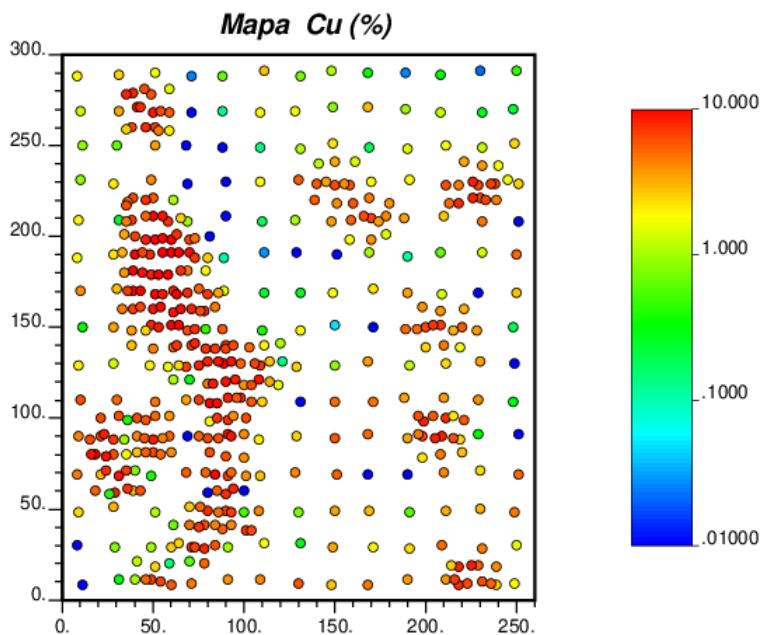


Figura C.4: Mapa de localização da variável cobre para o depósito do Walker Lake

## C.5 desagrupamento utilizando células móveis com o DECLUS

Para comparar as estatísticas interpoladas com as estatísticas dos dados é necessário utilizar alguma técnica de desagrupamento. O GSLIB permite realizar o desagrupamento a partir de céclulas móveis utilizando o programa DECLUS. Em seguida demonstramos o arquivo de parâmetros do programa.

```

Parameters for DECLUS
*****
START OF PARAMETERS:
Walker_Lake.txt      -file with data
1   2   0   3          - columns for X, Y, Z, and variable
-1.0e21    1.0e21     - trimming limits
declus.sum           -file for summary output

```

```

declus.out          -file for output with data & weights
1.0    1.0          -Y and Z cell anisotropy (Ysize=size*Yanis)
0                  -0=look for minimum declustered mean (1=max)
50   1.0  100        -number of cell sizes, min size, max size
5                  -number of origin offsets

```

Em seguida são demonstrados os significados dos parâmetros em cada linha do arquivo .par.

1. Endereço do arquivo do banco de dados
2. Colunas para a variável X, Y, Z e a variável desagrupada
3. Limites para o uso dos dados. Limite inferior ( $-10^{21}$ ) e limite superior ( $10^{21}$ )
4. Nome do arquivo de sumário estatístico do desagrupamento
5. Nome do arquivo de saída com os resultados e o peso de cada amostra
6. Anisotropia da célula na direção Y e em Z. (Tamanho em y = tamanho\*yanisotropia)
7. Procurar pela valor mínimo da média (0=mínimo, 1=máximo)
8. Número de células para o desagrupamento, os tamanhos mínimos e máximos da célula
9. Tamanho aumentado de cada célula. Essa medida evita com que valores na borda das células tenham problemas.

Podemos realizar o histograma dos dados desagrupados utilizando o programa HISTPLT. Para gerar o gráfico com os pesos basta apenas colocar na segunda linha a coluna dos pesos de desagrupamento do arquivo declus.out.

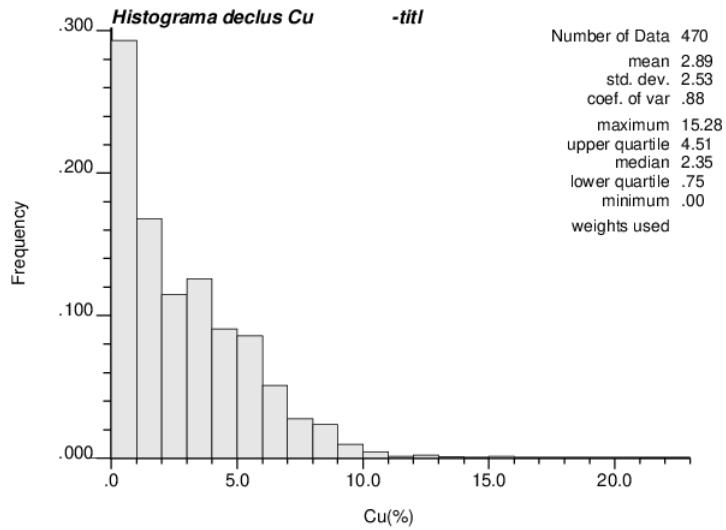


Figura C.5: Histograma dos dados desagrupados pelo GSLIB

## C.6 Convenção da orientação de eixos de anisotropia do GSLIB

O GSLIB possui uma convenção própria de orientação dos eixos para a modelagem variográfica. A direção principal da continuidade espacial é orientada segundo o eixo Y inicialmente. A figura C.6 demonstra este alinhamento dos eixos.

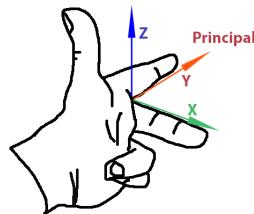


Figura C.6: Alinhamento inicial dos eixos do variograma segundo a notação do GSLIB

As rotações podem ser realizadas no sentido horário e anti-horário segundo este eixo de referência. A figura C.7 demonstra o sentido positivo e negativo de rotação dos eixos segundo a notação do GSLIB.

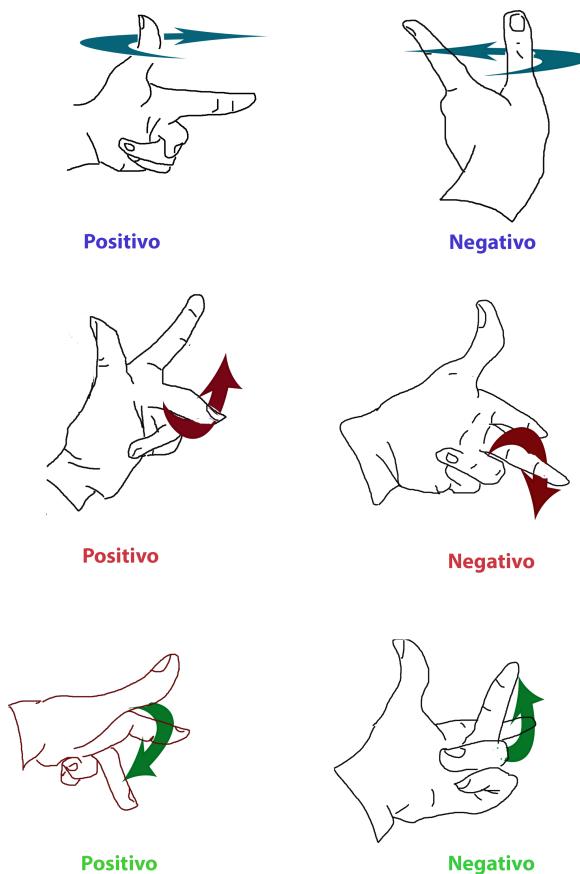


Figura C.7: Orientação dos eixos do variograma segundo a notação do GSLIB

## C.7 Variograma experimental (GAMV/ VARGPLT)

Variogramas experimentais podem ser facilmente criados com o programa GAMV. Além das funções tradicionais de variograma e covariograma, o programa também permite diversos outros tipos de funções como as funções pairwise, semimadograma e semivariograma de logaritmos. A seguir é demonstrado o arquivo .par para os variogramas experimentais de cobre do depósito Walker Lake.

```

Parameters for GAMV
*****
START OF PARAMETERS:
Walker_Lake.txt          -file with data
1   2   0                  -columns for X, Y, Z coordinates
1   3                  -number of variables,col numbers
-1.0e21      1.0e21      -trimming limits
Variograma_cobre.out      -file for variogram output

```

---

```

10                               -number of lags
10                               -lag separation distance
5                                -lag tolerance
2                                -number of directions
157.5  45 10    0.0   45   10  -azm,atol,bandh,dip,dtol,bandv
67.5   45 10    0.0   45   10  -azm,atol,bandh,dip,dtol,bandv
0                                -standardize sill? (0=no, 1=yes)
1                                -number of variograms
1      1     1                  -tail var., head var., variogram type

```

```

type 1 = traditional semivariogram
2 = traditional cross semivariogram
3 = covariance
4 = correlogram
5 = general relative semivariogram
6 = pairwise relative semivariogram
7 = semivariogram of logarithms
8 = semimadogram
9 = indicator semivariogram - continuous
10= indicator semivariogram - categorical

```

A seguir é demonstrado o significado do arquivo de parâmetros do GamV. É necessário que as informações descritas no arquivo de parâmetros sejam compatíveis, pois caso houver discrepâncias o programa envia uma mensagem de erro. O número de direções fornecidas, por exemplo, deve ser o mesmo do número de linhas contendo os parâmetros de azimute, mergulho, etc. O número de variáveis e o número das colunas relativas a cada uma destas variáveis também deve ser o mesmo.

1. Endereço do arquivo do banco de dados
2. Número da coluna X, Y e Z
3. Número de variáveis utilizadas e número das colunas
4. Limites para o uso dos dados. Limite inferior ( $-10^{21}$ ) e limite superior ( $10^{21}$ )
5. Número de lags utilizados no variograma (distância = número do lag\* tamanho do lag)
6. Tamanho do lag

7. Número de direções
8. Orientação : azimute, tolerância horizontal, banda horizontal, mergulho, tolerância vertical, banda vertical
9. Normalização do patarmar
10. Número de variogramas
11. variável da extermidade superior do vetor, variável da extremidade inferior do vetor e o tipo do variograma

Para realizar a plotagem do gráfico é necessário outro programa chamado de VARGPLT. Este programa gera o arquivo .ps contendo a imagem dos variogramas experimentais.

```

Parameters for VARGPLT
*****
START OF PARAMETERS:
vario_exp.ps           -file for PostScript output
2                      -number of variograms to plot
0.0      120            -distance limits (from data if max<min)
0.0      15              -variogram limits (from data if max<min)
0       9               -plot sill (0=no,1=yes), sill value)
Variograma experimental -Title for variogram
Variograma_cobre.out    -1 file with variogram data
1       1     1     1     1   - variogram #, dash #, pts?, line?, color
Variograma_cobre.out    -1 file with variogram data
2       2     1     1     7   - variogram #, dash #, pts?, line?, color

```

#### Color Codes for Variogram Lines/Points:

1=red, 2=orange, 3=yellow, 4=light green, 5=green, 6=light blue,  
 7=dark blue, 8=violet, 9=white, 10=black, 11=purple, 12=brown,  
 13=pink, 14=intermediate green, 15=gray

A seguir encontra-se a explicação de cada um dos parâmetros do arquivo. Lembre-se que antes de informar cada variograma é necessário colocar o arquivo de entrada do variograma experimental.

1. Nome do arquivo de saída PostScript (.ps)
2. Número de variogramas para plotagem (Um para cada direção e modelo)
3. Distância mínima e máxima do variograma experimental
4. Valor mínimo e máximo do variograma experimental
5. Plotagem do sill e o seu valor (0= não plotar, 1=plotar)
6. Título do variograma experimental
7. Arquivo contendo os dados do variograma experimental
8. Número do variograma, e parâmetros estéticos

O resultado dos variogramas podem ser vistos na figura C.8, onde a curva vermelha representa a direção de maior continuidade (Azimute = 157.5 graus) e a curva azul representa a direção de menor continuidade espacial (Azimute = 67.5 graus).

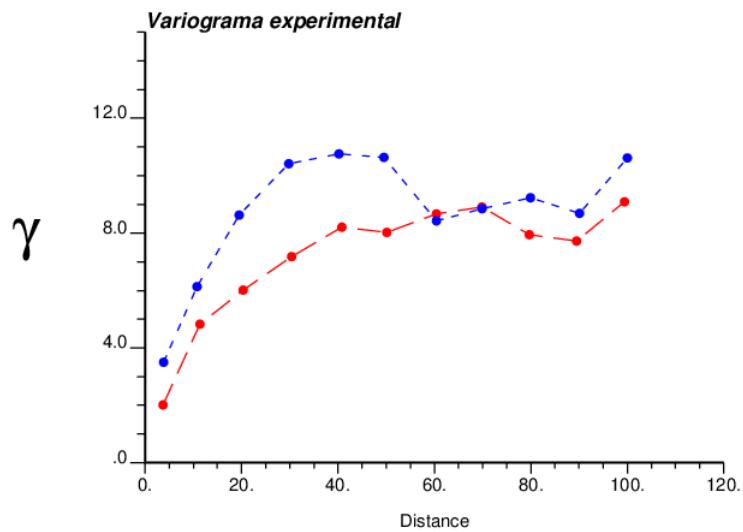


Figura C.8: Variogramas experimentais do Walker Lake. Curva vermelha representando a maior continuidade no azimute de 157.5 graus e curva preta representando a menor continuidade no azimute 67.5 graus norte.

## C.8 Modelagem de variogramas (VMODEL/VARGPLT)

Para criar modelos de variograma podemos utilizar o programa VMODEL para gerar um arquivo com a curva modelada. A seguir é demonstrado o arquivo de parâmetros do programa.

```

Parameters for VMODEL
*****
START OF PARAMETERS:
vmodel.var           -file for variogram output
2   100              -number of directions and lags
157.5   0.0   1      -azm, dip, lag distance
67.5    0.0   1      -azm, dip, lag distance
2     0.2              -nst, nugget effect
1     3   157.5  0.0   0.0  -it,cc,ang1,ang2,ang3
                  15   10   10.0  -a_hmax, a_hmin, a_vert
1     5.8  157.5  0.0   0.0  -it,cc,ang1,ang2,ang3
                  60   30   10.0  -a_hmax, a_hmin, a_vert

```

O GSLIB apresenta três tipos de modelos de continuidade espacial definidos por números. O tipo do variograma pode ser especificado de acordo com a seguinte lista

1. Variograma esférico
2. Variograma exponencial
3. Variograma gaussiano

A seguir encontra-se a explicação de cada um dos parâmetros do arquivo. Lembre-se que as informações fornecidas no arquivo devem ser compatíveis para que não ocorra erro na execução do programa. O número de direções fornecidas, por exemplo, deve ser o mesmo que o número de linhas contendo as informações do azimute e mergulho.

1. Nome do arquivo de saída do modelo de variograma
2. Número de direções e número de lags utilizados.
3. Para cada direção o azimute, mergulho e tamanho do lag
4. Número de estruturas e efeito pepita
5. Para cada estrutura o tipo do variograma, a contribuição, e os ângulos de rotação segundo a convenção GSLIB. (ang1 = rotação de Z, ang2 = rotação de X, ang3= rotação de Y). Para mais informações veja na seção C.6 a orientação dos eixos de anisotropia

6. Alcance na direção principal (orientada em Y inicialmente), alcance na direção mínima (orientada em X inicialmente) e alcance na direção vertical (orientada em Z inicialmente)

Para gerar o arquivo dos modelos de variograma juntamente com os dados experimentais, modificamos o arquivo VARGPLT anterior adicionando os dois variogramas a mais e modificando o número de variogramas plotados.

```

Parameters for VARGPLT
*****
START OF PARAMETERS:
vargplt.ps          -file for PostScript output
4                   -number of variograms to plot
0.0    120           -distance limits (from data if max<min)
0.0    15            -variogram limits (from data if max<min)
0      9             -plot sill (0=no,1=yes), sill value
Variograma experimental -Title for variogram
Variograma_cobre.out      -1 file with variogram data
1     1   1   1   1       - variogram #, dash #, pts?, line?, color
Variograma_cobre.out      -1 file with variogram data
2     2   1   1   7       - variogram #, dash #, pts?, line?, color
vmodel.var              -2 file with variogram data
1     0   0   1   6       - variogram #, dash #, pts?, line?, color
vmodel.var              -2 file with variogram data
2     0   0   1   6 -  variogram #, dash #, pts?, line?, color

```

#### Color Codes for Variogram Lines/Points:

1=red, 2=orange, 3=yellow, 4=light green, 5=green, 6=light blue,  
 7=dark blue, 8=violet, 9=white, 10=black, 11=purple, 12=brown,  
 13=pink, 14=intermediate green, 15=gray

A figura C.9 representa o ajuste do variograma para a variável cobre no depósito Walker Lake.

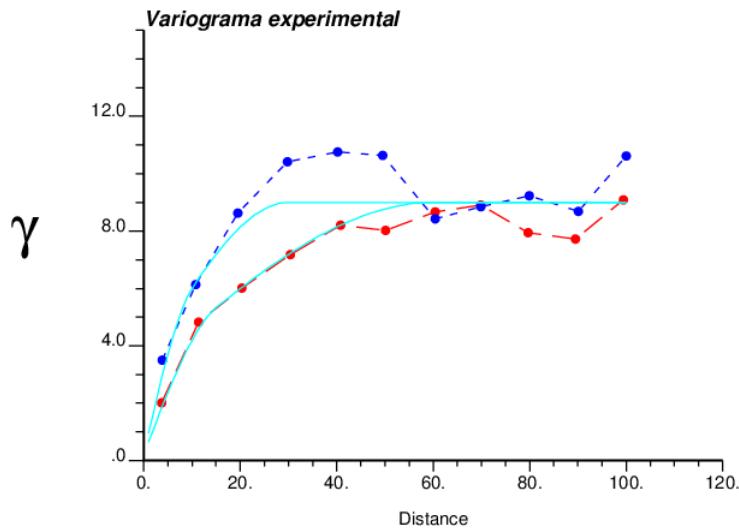


Figura C.9: Modelo de variograma com duas estruturas para o Walker Lake. Curva vermelha representando a maior continuidade no azimute de 157.5 graus e curva preta representando a menor continuidade no azimute 67.5 graus norte. Curva azul claro representado o modelo.

## C.9 Validação Cruzada com (KT3D/LOCMAP)

Para ajustar a melhor estratégia de busca das amostras na krigagem, bem como também ajustar o melhor modelo de variograma, podemos utilizar a validação cruzada como uma alternativa para identificar os possíveis erros de estimativa. A seguir é demonstrado o arquivo de parâmetros do programa KT3D utilizado não apenas para a validação cruzada, mas também para a krigagem. O programa KT3D não apenas realiza krigagem ordinária, mas diversos outros tipos de krigagem. Neste exemplo realizaremos apenas a krigagem ordinária. Em dados tridimensionais uma alternativa mais interessante que a validação cruzada é o jackknife, em que furos são retirados como um todo e os erros computados em cada um deles. Como nos casos tridimensionais as compostas são muito próximas uma das outras, os erros de validação cruzada tendem a ser muito pequenos, falseando um melhor ajuste dos dados.

```

Parameters for KT3D
*****
START OF PARAMETERS:
Walker_Lake.txt          -file with data
0 1 2 0 3 0               - columns for DH,X,Y,Z,var,sec var

```

```

-1.0e21    1.0e21          - trimming limits
1           -option: 0=grid, 1=cross, 2=jackknife
xvk.dat      -file with jackknife data
1   2   0   3   0          - columns for X,Y,Z,vr and sec var
3           -debugging level: 0,1,2,3
kt3d.dbg      -file for debugging output
kt3d.out      -file for kriged output
25  0.1   10             -nx,xmn,xsiz
30  0.1   10             -ny,ymn,ysiz
1   0.0   1.0            -nz,zmn,zsiz
1   1     1              -x,y and z block discretization
2   3                 -min, max data for kriging
2                   -max per octant (0-> not used)
40.0 30.0 10.0         -maximum search radii
157.5 0.0  0.0          -angles for search ellipsoid
0   2.302             -0=SK,1=OK,2=non-st SK,3=exdrift
0 0 0 0 0 0 0 0 0       -drift: x,y,z,xx,yy,zz,xy,xz,zy
0           -0, variable; 1, estimate trend
extdrift.dat        -gridded file with drift/mean
4           - column number in gridded file
2   0.2               -nst, nugget effect
1   3 157.5 0 0   -it, c, azm, a_max, a_min
                  -a_hmax, a_hmin, a_vert
1   5.8 157.5 0 0   -it, c, azm, a_max, a_min
                  -a_hmax, a_hmin, a_vert

```

A seguir encontra-se a explicação de cada um dos parâmetros do arquivo.

1. Arquivo contendo os dados
2. colunas para o índice do furo, coordenadas X,Y,Z a variável de interesse e uma variável secundária
3. Limites para o uso dos dados. Limite inferior ( $-10^{21}$ ) e limite superior ( $10^{21}$ )
4. Opções da krigagem (0= em malha, 1= validação cruzada, 2= jackknife). Marque 1 para realizar a validação cruzada.
5. arquivo para utilização do jackknife. Caso não exista este arquivo o programa simplesmente ignora esta instrução.

6. Variável X, Y, Z a variável de interesse e a variável secundária. Adicione 0 caso não possua esta informação em algum tópico. Por exemplo em casos 2D utilizamos a coordenada Z = 0.
7. Nível de depuração da krigagem. Quanto maior o nível de depuração, maiores serão as informações contidas no arquivo de depuração .dgb. Para um nível 3 de depuração é possível verificar as matrizes de krigagem e dados utilizados no cálculo.
8. Nome do arquivo de depuração
9. Nome do arquivo de saída da krigagem
10. número de células em X, o valor mínimo do grid em X e o tamanho da célula em X
11. número de células em Y, o valor mínimo do grid em Y e o tamanho da célula em Y
12. número de células em Z, o valor mínimo do grid em Z e o tamanho da célula em Z
13. Discretização dos blocos da krigagem em x, y e z para krigagem de blocos.
14. Número mínimo de dados utilizados e número máximo.
15. Máximo número de dados por octante. (0 para não ser utilizado)
16. Tamanho dos eixos de anisotropia para a estratégia de busca. Tamanho na direção principal, tamanho na direção mínima e tamanho na direção vertical.
17. Ângulos de rotação do elipsóide de anisotropia segundo a convenção GSLIB. Rotação no eixo Z, rotação no eixo X e rotação no eixo Y. Verificar na seção [C.6](#) a conveção do programa.
18. Escolha do tipo de krigagem (0 = krigagem simples, 1= krigagem ordinária, 2= krigagem não estacionária, 3= krigagem com tendência externa) e o valor médio caso seja selecionada a krigagem simples. Em outros casos o programa ignora este parâmetro.
19. Coeficiente dos polinômios em caso de krigagem não estacionária.
20. Estimativa da variável ou da tendência (0= variável, 1= tendência)

21. Arquivo com tendência externa caso seja realizada krigagem com tendência externa.
22. Coluna da tendência no arquivo
23. Adição dos variogramas: Número de estruturas e efeito pepita
24. Para cada estrutura p tipo do modelo de variograma, contribuição, rotação no eixo Z, rotação no eixo X e rotação no eixo Y segundo a referência do GSLIB.
25. Para cada estrutura os alcances na direção principal, mínima e vertical.

Para a plotagem do mapa de validação cruzada podemos utilizar o arquivo LOCMAP lembrando sempre de marcar a opção de validação cruzada. A seguir é demonstrado o arquivo de parâmetros .par do programa LOCMAP.

```
Parameters for LOCMAP
*****
START OF PARAMETERS:
kt3d.out          -file with data
1    2    7          - columns for X, Y, variable
-1.0   1.0e21       - trimming limits
cv_map.ps         -file for PostScript output
0.0    250          -xmn, xmx
0.0    300.          -ymn, ymx
1                  -0=data values, 1=cross validation
1                  -0=arithmetic, 1=log scaling
1                  -0=gray scale, 1=color scale
0                  -0=no labels, 1=label each location
0.01   10.0   1.      -gray/color scale: min, max, increm
0.5                -label size: 0.1(sml)-1(reg)-10(big)
Validacao Cruzada -Title
```

A figura C.10 demonstra o mapa da validação cruzada para a variável cobre no depósito Walker Lake.

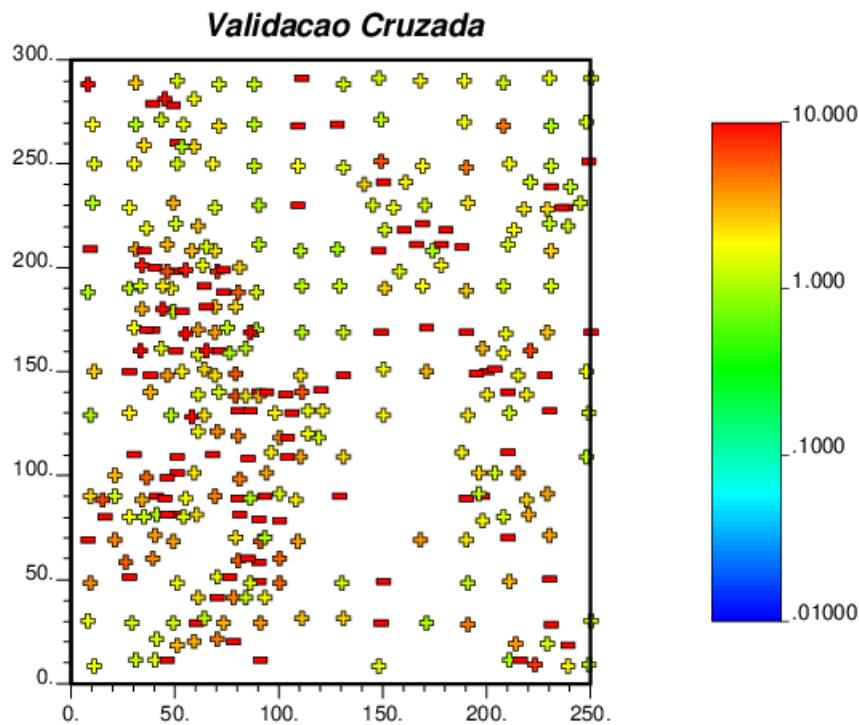


Figura C.10: Mapa da validação cruzada para a variável cobre, no depósito Walker Lake.

## C.10 Krigagem com (KT3D/PIXELPLT)

O arquivo de parâmetros para a krigagem é o mesmo que o utilizado para a validação cruzada, apenas alterando a opção de validação cruzada (1) para malha (0). A explicação dos parâmetros do arquivo pode ser obtida na seção anterior. A seguir é demonstrado o arquivo do tipo .par utilizado para a krigagem da variável cobre no depósito Walker Lake.

```

Parameters for KT3D
*****
START OF PARAMETERS:
Walker_Lake.txt          -file with data
0 1 2 0 3 0               - columns for DH,X,Y,Z,var,sec var
-1.0e21  1.0e21           - trimming limits
0                           -option: 0=grid, 1=cross, 2=jackknife
xvk.dat                    -file with jackknife data
1 2 0 3 0                 - columns for X,Y,Z,vr and sec var

```

```

3                               -debugging level: 0,1,2,3
kt3d.dbg
kt3d.out
25  0.1    10                 -nx,xmn,xsiz
30  0.1    10                 -ny,ymn,ysiz
1   0.0    1.0                -nz,zmn,zsiz
1   1      1                  -x,y and z block discretization
1   10
0
80  60 10.0                  -min, max data for kriging
0                           -max per octant (0-> not used)
157.5   0.0    0.0            -maximum search radii
0      2.302
0 0 0 0 0 0 0 0 0 0          -angles for search ellipsoid
0
extdrift.dat
4
2   0.2                      -nst, nugget effect
1   3  157.5  0  0  -it, c, azm, a_max, a_min
           15  10.0  1      -a_hmax, a_hmin, a_vert
1   5.8 157.5  0  0  -it, c, azm, a_max, a_min
       60  30   1      -a_hmax, a_hmin, a_vert

```

Para a plotagem do mapa de krigagem utilizamos o programa PIXELPLT. A seguir é demonstrado o arquivo .par do programa.

```

Parameters for PIXELPLT
*****
START OF PARAMETERS:
kt3d.out
1
-1.0e21  1.0e21
krig_interp.ps
1
25  0.1    1.0
30  0.1    1.0
1   1      1.0
1
1
Krigagem Cobre
Leste

```

-file with gridded data  
 - column number for variable  
 - data trimming limits  
 -file with PostScript output  
 -realization number  
 -nx,xmn,xsiz  
 -ny,ymn,ysiz  
 -nz,zmn,zsiz  
 -slice orientation: 1=XY, 2=XZ, 3=YZ  
 -slice number  
 -Title  
 -X label

```

Norte                                -Y label
0                                     -0=arithmetic, 1=log scaling
1                                     -0=gray scale, 1=color scale
0                                     -0=continuous, 1=categorical
0.0 20.0 1.0                         -continuous: min, max, increm.
4                                     -categorical: number of categories
1     3     Code_One                  -category(), code(), name()
2     1     Code_Two
3     6     Code_Three
4     10    Code_Four

```

Color Codes for Categorical Variable Plotting:

1=red, 2=orange, 3=yellow, 4=light green, 5=green, 6=light blue,  
 7=dark blue, 8=violet, 9=white, 10=black, 11=purple, 12=brown,  
 13=pink, 14=intermediate green, 15=gray

A explicação dos parâmetros do arquivo é demonstrada a seguir

1. Nome do arquivo dos dados krigagdos
2. Coluna da variável a ser plotada
3. Limites para o uso dos dados. Limite inferior ( $-10^{21}$ ) e limite superior ( $10^{21}$ )
4. Nome do arquivo de saída .ps
5. Número de realizações. No caso de krigagem possuímos apenas uma realização.  
 No entanto nos casos de simulação podem haver mais de uma realização.
6. Número de células em X, valor mínimo da coordenada X e tamanho da célula e X
7. Número de células em Y, valor mínimo da coordenada Y e tamanho da célula e Y
8. Número de células em Z, valor mínimo da coordenada Z e tamanho da célula e Z
9. Orientação do corte realizado em malhas tridimensionais. (1=orientação horizontal XY, 2= Corte vertical em XZ, 3= Corte vertical em YZ)
10. Número do corte

11. Título do gráfico
12. Título da coordenada leste
13. Título da coordenada Norte
14. Escala aritmética ou logarítmica (0=aritmética, 1=logarítmica)
15. Escala de cinza ou de cores (0=cinza, 1=de cores)
16. Variável de entrada contínua ou categórica (0=contínua, 1=categórica)
17. Número mínimo, máximo e incremento da variável plotada
18. Número de categorias se selecionada variável categórica
19. Categoria, cor e nome da categoria

As figuras C.11 e C.12 demonstram os mapas krigados e a variância de krigagem para o depósito Walker Lake.

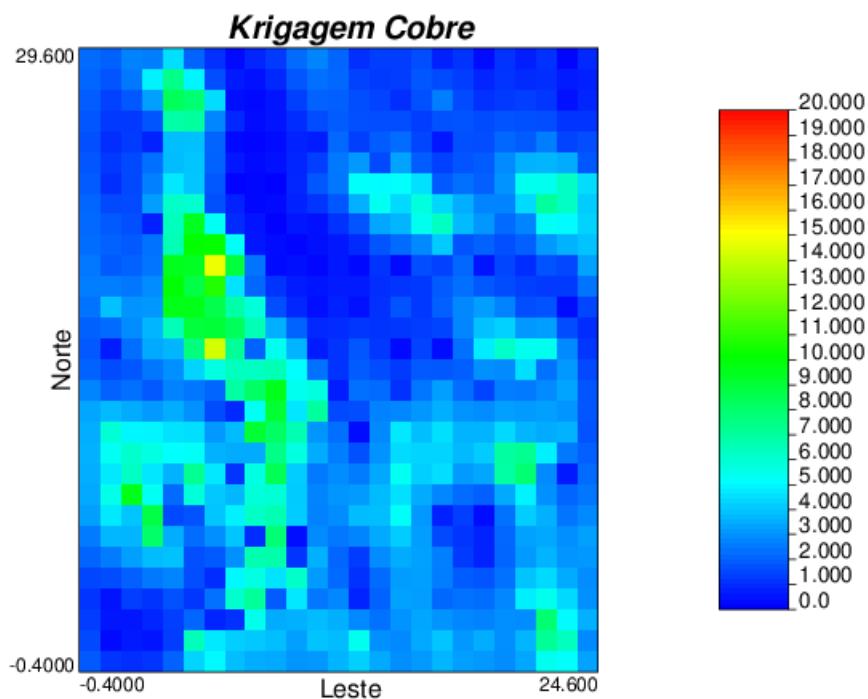


Figura C.11: Mapa da krigagem para a variável cobre, no depósito Walker Lake.

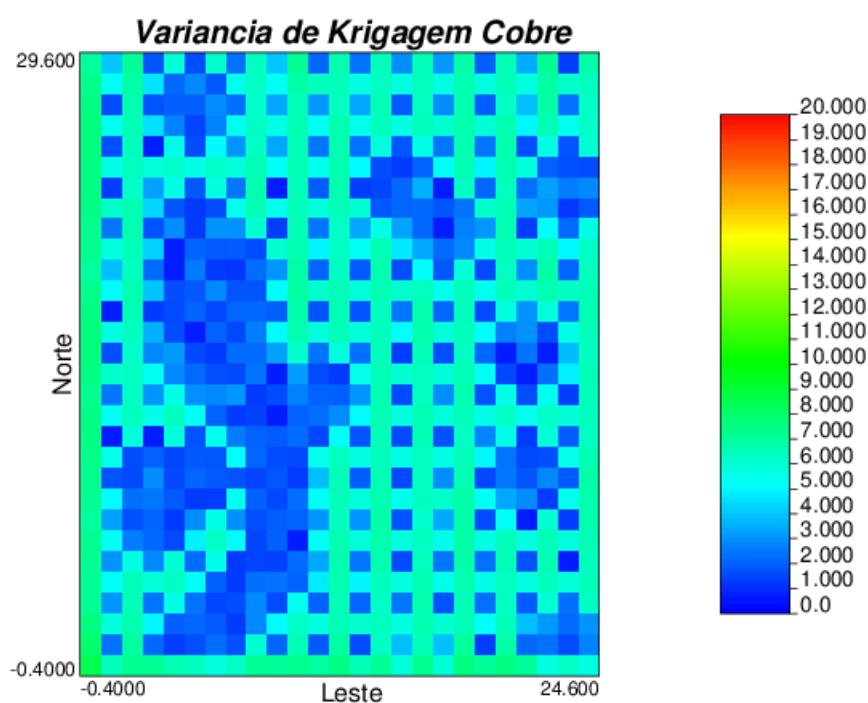


Figura C.12: Mapa da variância de krigagem para a variável cobre, no depósito Walker Lake.



## D. Geoestatística utilizando o SGeMS

O SGeMS (Stanford Geostatistical Modeling Software) é um programa de livre distribuição para a solução de problemas relacionados às variáveis espacialmente relacionadas. Diferentemente dos demais softwares gratuitos, o SGeMS possui interface amigável e de fácil utilização. O software pode ser encontrado no seguinte link <http://sgems.sourceforge.net/?q=node/77>, e baixado nas suas respectivas versões 32bits e 64bits. A figura D.1 demonstra a interface principal do programa.

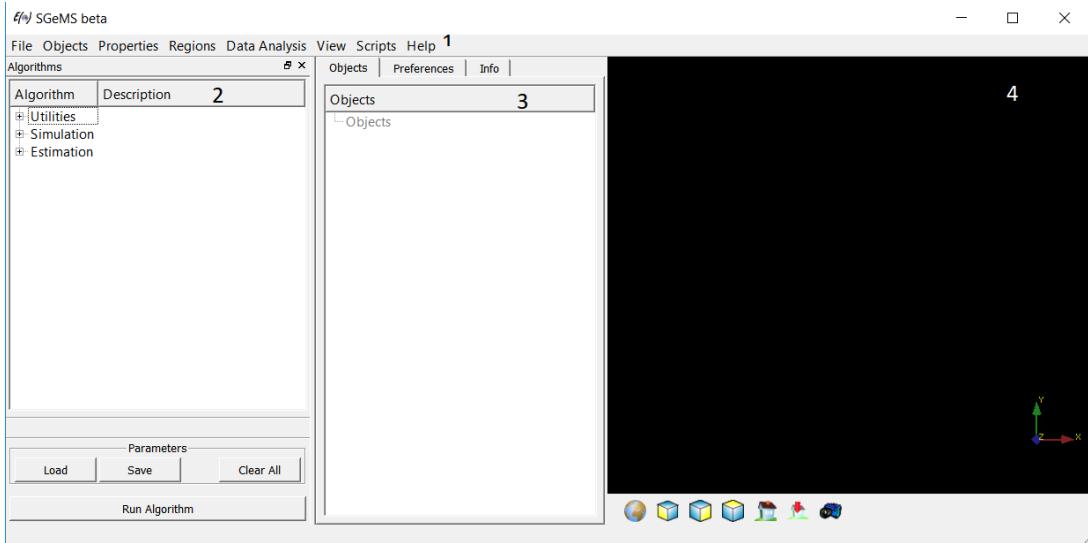


Figura D.1: Janela do SGeMs principal. 1) Menu push down contendo principais informações do programa como importação de arquivos, realização de estatísticas e geração de scripts. 2) Aba de algoritmos contendo os principais algoritmos do SGeMS 3) Objetos importados para o SGeMs como arquivos de pontos ou grids 4) Visualização das informações espacialmente.

## D.1 Importando um arquivo de pontos no SGeMS

Para importar um arquivo de pontos no SGeMS precisamos recorrer ao menu pull down na aba objetos, e em seguida clicar em Load Data. Os arquivos de importação para o SGeMS seguem a mesma referência dos arquivos ASCII importados pelo GSLib, contendo inicialmente o nome do arquivo, número de colunas, nome das variáveis e os dados. Mais informações revisar a seção C.3. A figura D.2 demonstra o procedimento de entrada.

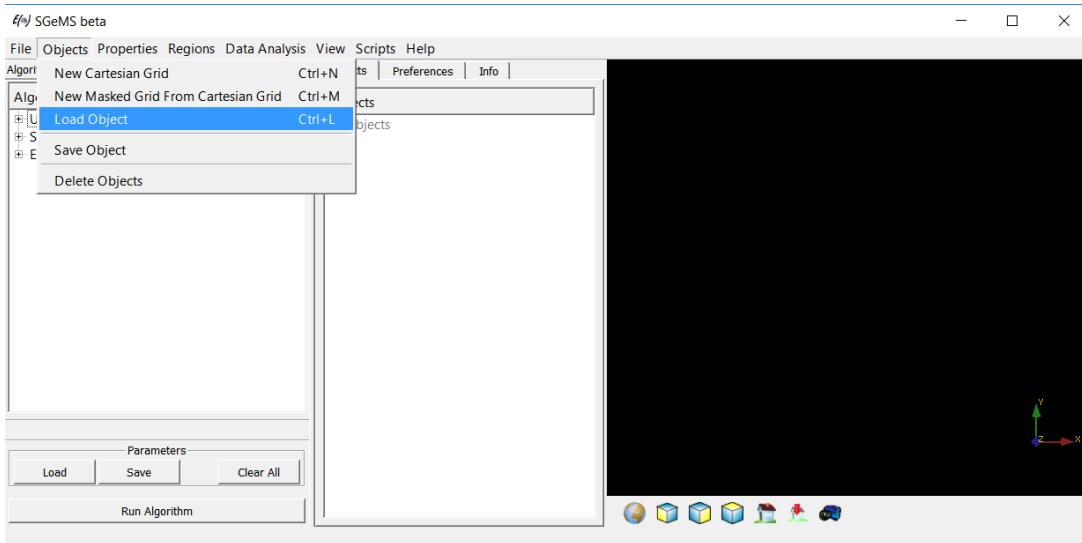


Figura D.2: Importação de arquivo de pontos no SGeMs. Seleção no menu pull down.

Em seguida é aberta uma nova aba para a seleção do arquivo de pontos. Ao escolher o endereço do arquivo é requisitado se o tipo de importação ocorrerá para um arquivo de pontos ou de grid. Neste caso desejamos que o arquivo selecionado seja de pontos. A figura D.3 demonstra a importação dos dados.

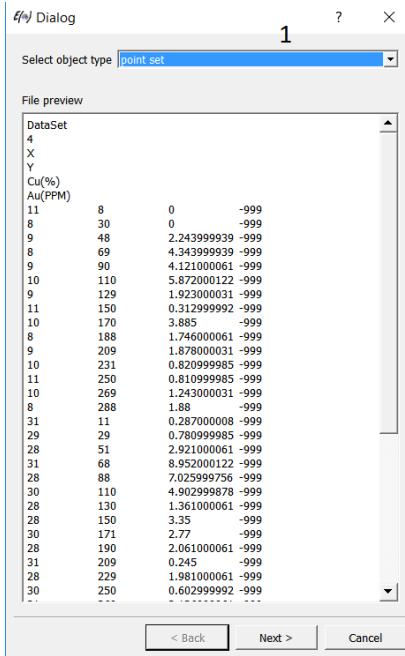


Figura D.3: Importação de arquivo de pontos no SGeMs. Definição do arquivo de grid. 1) Determinação do arquivo de grids e de pontos.

Uma nova aba é aberta possibilitando criar o objeto de pontos no SGeMs. A figura D.4 demonstra os parâmetros necessários para se criar este arquivo. É necessário atribuir um nome ao objeto, definir quais variáveis são caracterizadas pelas coordenadas espaciais e o número relacionado aos dados faltantes no banco de dados.

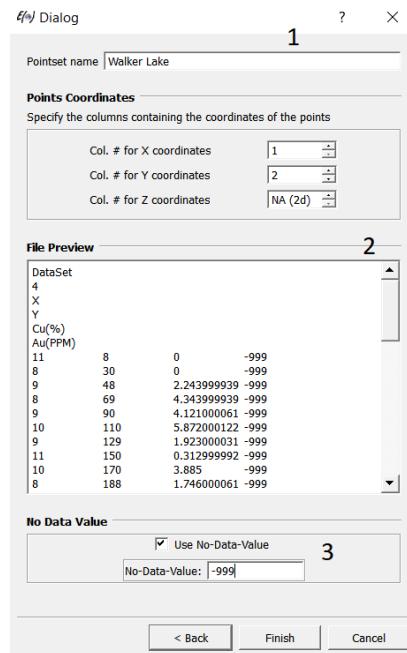


Figura D.4: Importação de arquivo de pontos no SGeMs. 1) Denominação do objeto relacionado ao arquivo de pontos no SGeMs 2) Atribuição das variáveis das coordenadas no banco de dados 3) Definição do valor considerado como dado faltante

## D.2 Visualização dos dados - Mapa de localização

Para visualizar o arquivo de pontos no SGeMs basta selecionar na aba de objetos a propriedade desejada do banco de dados. Ao lado direito, na aba de visualizações, poderá ser observado as amostras com seus respectivos valores coloridos no mapa.

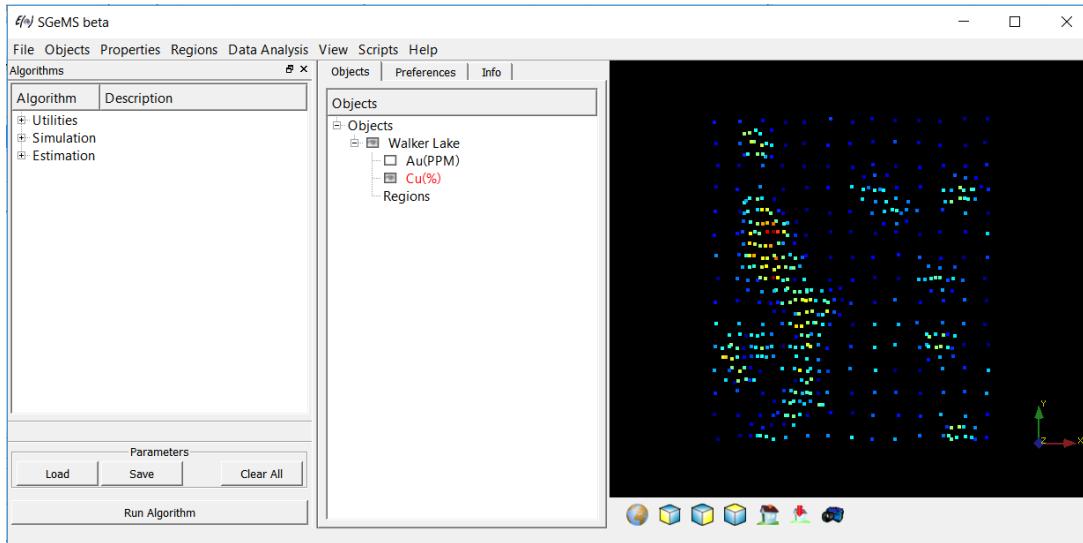


Figura D.5: Mapa de localização da variável Cobre - Seleção da propriedade no menu de objetos

### D.3 Criação do histograma

Para criar um histograma basta clicar na propriedade considerada com o botão direito e selecionar a opção histograma. A figura D.6 demonstra os passos necessários para a geração do histograma.

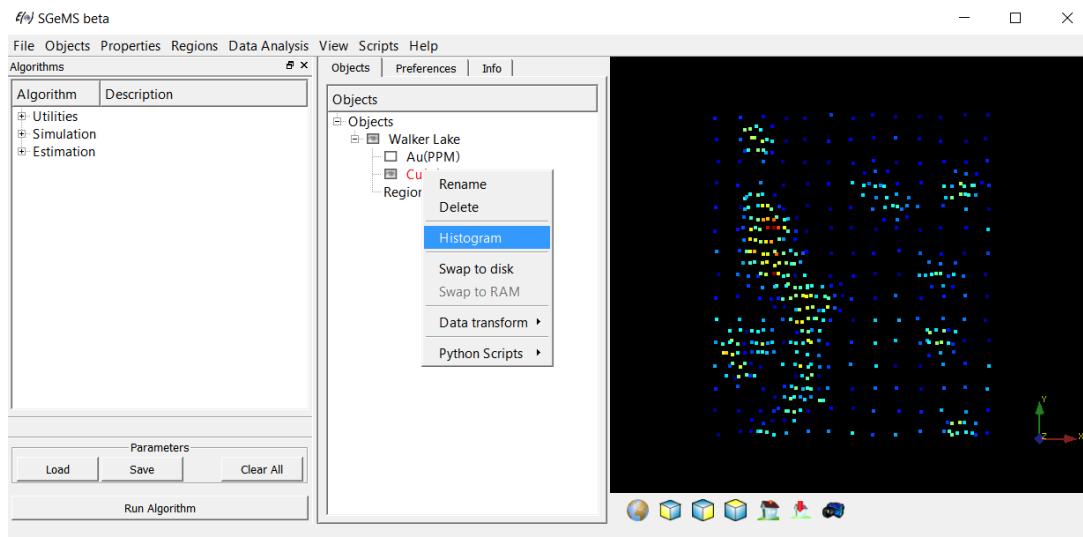


Figura D.6: Histograma da variável cobre - Seleção pela aba objeto

Em seguida é demonstrada a

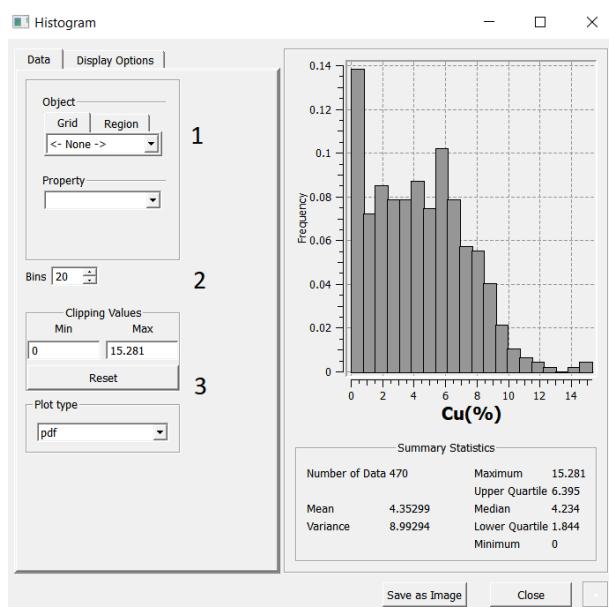


Figura D.7: Histograma da variável cobre - Determinação da propriedade e parâmetros. 1) Determinação





# Bibliografia

- M. S. d. Alencar. Teoria de conjuntos, medida e probabilidade. *São Paulo: Ed. Érica*, 2014.
- G. J. Borradaile. *Statistics of earth science data: their distribution in time, space and orientation*. Springer Science & Business Media, 2013.
- K. E. Brassel and D. Reif. A procedure to generate thiessen polygons. *Geographical Analysis*, 11(3):289–303, 1979.
- P. Carrasco, J.-P. Chilès, and S. A. Séguret. Additivity, metallurgical recovery, and grade. In *8th international Geostatistics Congress*, pages on–CD, 2008.
- G. Casella and R. L. Berger. Inferência estatística. *São Paulo: Cengage Learning*, 2010.
- J.-P. Chiles and P. Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons, 2009.
- J.-P. Delhomme. Applications de la théorie des variables régionalisées dans les sciences de l'eau. 1978.
- J. DeutschCV. Gslib: Geostatistical software libraryanduser'sguide., 1998.
- H. d. A. FEITOSA, M. d. NASCIMENTO, and A. BRUNO-ALFONSO. Teoria dos conjuntos: sobre a fundamentaçao matemática e a construçao de conjuntos numéricos. *Rio de Janeiro: Editora Ciêncie Moderna*, 2011.

- M. M. C. Figueira. Identificação de outliers. *Millenium*, 1998.
- P. Goovaerts. *Geostatistics for natural resources evaluation*. Oxford University Press on Demand, 1997.
- N. Gustavsson, E. Lampio, and T. Tarvainen. Visualization of geochemical data on maps at the geological survey of finland. *Journal of Geochemical Exploration*, 59 (3):197–207, 1997.
- P. Gy. *Sampling of particulate materials theory and practice*, volume 6. Elsevier, 2012.
- F. Heylighen. Occam’s razor. *Principia cybernetica web*, 1997.
- W. Hustrulid, M. Kuchta, and R. Martin. Open pit mine planning and design, volume 1: Fundamentals. *CRC Press Taylor & Francis Group*, 6000:33487–2742, 2006.
- E. H. Isaaks and R. M. Srivastava. Applied geostatistics. 1989.
- A. G. Journel and C. J. Huijbregts. *Mining geostatistics*. Academic press, 1978.
- D. Krige. On the departure of ore value distributions from the lognormal model in south african gold mines. *Journal of the Southern African Institute of Mining and Metallurgy*, 61(4):231–244, 1960.
- R. d. S. Machado. Uma alternativa para a estimativa de teores em depósitos de ouro: Geoestatística paramétrica de campo. 2012.
- R. J. L. Maranhao. *Introdução à pesquisa mineral*. Banco do Nordeste do Brasil SA Escritório Técnico de Estudos Econômicos do . . . , 1985.
- G. Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.
- C. J. Morris, D. S. Ebert, and P. L. Rheingans. Experimental analysis of the effectiveness of features in chernoff faces. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*, volume 3905, pages 12–17. International Society for Optics and Photonics, 2000.
- M. Pyrcz and C. Deutsch. Declustering and debiasing. *Newsletter*, 19, 2003.
- J.-M. Rendu. *An introduction to cut-off grade estimation*. Society for Mining, Metallurgy, and Exploration, 2014.

- M. E. Rossi and C. V. Deutsch. *Mineral resource estimation*. Springer Science & Business Media, 2013.
- R. M. Srivastava and H. M. Parker. Robust measures of spatial continuity. In *Geostatistics*, pages 295–308. Springer, 1989.
- A. Susaeta, E. Rubio, G. Pais, and J. Enriquez. Dilution behaviour at codelco panel cave mines. *Proceedings of MassMin 2008*, pages 167–178, 2008.
- S. Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.
- J. K. Yamamoto. *Avaliação e classificação de reservas minerais*, volume 38. Edusp, 2001.

