Ministério de Minas e Energia
Secretaria de Geologia, Mineração e Transformação Mineral
Serviço Geológico do Brasil - CPRM
Diretoria de Relações Institucionais e Desenvolvimento

INTERPRETAÇÃO DE DADOS DE PROSPECÇÃO GEOQUÍMICA COM O AUXÍLIO DE ESTATÍSTICA

José Leonardo Silva Andriotti

Superintendência Regional de Porto Alegre
Agosto de 2010

MINISTÉRIO DE MINAS E ENERGIA
Márcio Pereira Zimmermann
Ministro de Estado

SECRETARIA DE GEOLOGIA, MINERAÇÃO E TRANSFORMAÇÃO MINERAL
Cláudio Scliar
Secretário

COMPANHIA DE PESQUISA DE RECURSOS MINERAIS - CPRM
Serviço Geológico do Brasil

Agamenon Sérgio Lucas Dantas
Diretor-Presidente

Manoel Barretto da Rocha Neto
Diretor de Geologia e Recursos Minerais

José Ribeiro Mendes
Diretor de Hidrologia e Gestão Territorial

Fernando Pereira de Carvalho
Diretor de Relações Institucionais e Desenvolvimento

Eduardo Santa Helena
Diretor de Administração e Finanças

Sabino Orlando Conceição Loguércio
Chefe do Departamento de Apoio Técnico - DEPAT

José Alcides Fonseca Ferreira
Superintendente Regional de Porto Alegre - SUREG/PA

José Leonardo Silva Andriotti
Gerente de Relações Institucionais e Desenvolvimento - GERIDE/PA

Ministério de Minas e Energia
Secretaria de Geologia, Mineração e Transformação Mineral
Serviço Geológico do Brasil - CPRM
Diretoria de Relações Institucionais e Desenvolvimento

INTERPRETAÇÃO DE DADOS DE PROSPECÇÃO GEOQUÍMICA COM O AUXÍLIO DE ESTATÍSTICA

José Leonardo Silva Andriotti

Superintendência Regional de Porto Alegre
Agosto de 2010

Foto da capa: campanha de prospecção geoquímica (arquivo da CPRM)

Ficha Catalográfica

A573 Andriotti, José Leonardo Silva

Interpretação de dados de prospecção geoquímica com o auxílio de estatística / José Leonardo Silva Andriotti. – Porto Alegre: CPRM, 2010.

73 p. : ilust.

1. Prospecção Geoquímica 2. Estatística 3. Geoestatística I. Título.

CDU - 550.41:519.5

CPRM – Superintendência Regional de Porto Alegre

Gerência de Relações Institucionais e Desenvolvimento – GERIDE

Biblioteca Regional de Porto Alegre

Bibl. Ana Lúcia B. F. Coelho – CRB10/840

SUMÁRIO

LISTA DE ILUSTRAÇÕES

FIGURAS

TABELAS

INTRODUÇÃO

No presente trabalho são tratados alguns temas e discutidas algumas técnicas de tratamento de dados utilizadas em prospecção geoquímica, aí incluídos procedimentos de campo e de laboratório. Não são discutidas todas as técnicas e procedimentos usuais, uma vez que algumas abordagens, como a Geoestatística, por exemplo, são discutidas em detalhes em outros documentos do mesmo autor. Procurou-se apresentar de forma resumida técnicas que, no entendimento do autor, devem ser conhecidas por quem trabalha com dados analíticos gerados por campanhas de prospecção geoquímica.

O entendimento dos conceitos aqui tratados pressupõe uma certa familiaridade do leitor com alguns conceitos de base da Estatística, como as noções de população e amostra, alguns tipos de distribuições mais usuais em geoquímica (normal, lognormal), parâmetros de tendência central e de dispersão e inferência estatística, bem como familiaridade com os conceitos fundamentais da Geoquímica. Na Estatística Multivariada a concentração se dá nas abordagens mais em uso nos últimos tempos, como as análises fatorial e de componentes principais e a análise de agrupamentos (outras técnicas, como os testes de hipótese multivariados, a análise discriminante e a análise de variância multivariada não foram incluídas não por terem sido consideradas como de menor utilidade, mas por não ter sido possível incluí-las na programação de espaço e tempo a que se propõe este documento, que é de caráter introdutório).

Críticas e sugestões sobre o conteúdo deste documento serão bem recebidos pelos mails *andri@portoweb.com.br* ou *jose.andriotti@cprm.gov.br*.

PROSPECÇÃO GEOQUÍMICA E ESTATÍSTICA

Um problema sempre presente na interpretação de dados gerados em campanhas de prospecção geoquímica (em seus vários meios de amostragem) é a determinação dos limiares, valores que separam as faixas de valores definidas como representando o *background* e a de valores anômalos. No decorrer das últimas décadas este assunto cresceu de importância nas discussões entre os geoquímicos, o que faz sentido por ser a determinação de anomalias de fundamental importância, pois com base nas áreas definidas ou classificadas como tal são definidas etapas seguintes da exploração. A realidade é única, e diferentes métodos tentam nos dizer qual é (e como é) ela ou como ela se apresenta, de diferentes formas, às vezes até com alterações significativas em quantidades de amostras, em valores, em zonas delimitadas. No presente trabalho são apresentados alguns dos métodos mais utilizados e comentados, o assunto não se esgota por ser tema que tem gerado novidades nos últimos anos.

Uma questão importante a se levar em conta quando se discute sobre anomalias geoquímicas é que não sabemos antecipadamente se elas existem no conjunto de dados disponível, e outra questão importante é sua localização geográfica. Nas discussões aqui tratadas não são consideradas situações que envolvem mais diretamente o entendimento de campo do que o tratamento da informação numérica, como, por exemplo, anomalias (ou faixas de valores) de solo deslocadas por se encontrarem em encostas, ou seja, vamos discutir as determinações de valores considerando que as questões de campo já foram devidamente compreendidas e solucionadas. É sabido que todo tratamento de informação se baseia no fato de que a amostragem foi executada de forma correta, sem erros ou mistura de populações. Campanhas regionais de geoquímica de sedimentos de corrente que drenam diferentes litologias têm seus resultados tratados, conforme muitos casos relatados na bibliografia sobre o assunto, considerando um único *background* para fins estatísticos de tratamento, o que é passível de ser aceito se for levado em conta que a mistura de material durante o transporte homogeinize as variabilidades geradas pelas diferentes fontes, no caso as diferentes litologias drenadas. Outro assunto a ser considerado, quando se trata de sedimentos de drenagem (cujas bacias de captação têm tamanhos diferentes em área) é a representação dos resultados, neste caso mapas de contorno não são a melhor opção, sendo indicados mapas com círculos com diâmetros (ou símbolos, ou cores) variando em função do valor numérico do atributo representado.

Técnicas de determinação de *threshold* baseadas em modelos estatísticos de distribuição muitas vezes se mostram pouco eficazes. Anomalia e *background* são respostas a processos distintos, cada um representado por seu respectivo histograma. Em uma área em que afloram rochas calcárias e ultramáficas e a variável de estudo for, por exemplo, o Cr, que tem valores

médios respectivos significativamente diferentes, a não separação poderá mascarar a presença de eventuais mineralizações em uma destas litologias, e os parâmetros estatísticos não representarão a situação existente da melhor forma.

Muitos autores têm tratado do tema de determinação de anomalias em prospecção geoquímica nas últimas décadas, entre eles podem ser citados Parslow (1974), Sinding-Larsen (1977), Miesch (1981), Howart (1983), Sinclair (1986, 1991), Velasco & Verma (1998), Reinmann & Filzmozer (2000), Matschullat et al. (2000), Karger & Sandomirsky (2001), Bounessah (2003), Rantitsch (2004), Inacio et al. (2004), Filzmozer et al. (2005) e Carranza (2009), além de outros citados pontualmente dentro do presente trabalho em situações específicas. Uma revisão com ênfase a procedimentos de laboratório foi publicada por Andriotti (2005).

Diferentes autores sugerem diferentes métodos para encontrar o valor de limiar. Hawkes e Webb (1962) sugerem o uso de duas vezes o valor do *background* como *threshold*. Bolviken (1971) e Tennant e White (1959) mostraram que não há um valor simples de *threshold* mas uma distribuição de valores de *background* e uma distribuição de valores anômalos. Rose e colaboradores (1979) usam regressão múltipla para estimar a concentração de *background*.

Stanley a Sinclair (1989) propõem uma classificação de técnicas de seleção de *thresholds* em três categorias principais. A primeira é baseada em *métodos experimentais*, que dependem da experiência dos técnicos e utilizam tabulações de dados ou avaliação visual de histogramas, ressaltando os valores mais elevados de modo subjetivo. Não são adequados onde é importante a classificação de indivíduos como anômalos ou como pertencentes à população de *background* (caso de trabalhos regionais). A segunda pode ser classificada como abordagens baseadas em *modelos subjetivos*, aplicam algum tipo de modelo matemático ou estatístico aos dados geoquímicos. A terceira classe é a das abordagens baseadas em *modelos objetivos*, que diferem das anteriores apenas no sentido de que os *thresholds* são definidos com base nos próprios dados em vez de sobre decisões arbitrárias dos técnicos de exploração. Dois exemplos desta classe são os gráficos de probabilidade (Sinclair, 1974 e 1976) e a estatística GAP (Miesch, 1981), esta última teve uso restrito.

Segundo muitos autores não há uma boa razão para continuar usando (média $\pm$ 2 desvios padrão) como regra geral para determinação de limiares geoquímicos, originalmente proposta como uma forma para identificar 2,5% dos dados no extremo superior da curva. O argumento é o de que não há uma razão técnica para que se decida antecipadamente que existem 2,5% de valores significativamente elevados e que não existem, por exemplo, 4%, 5%

ou 10% de valores significativamente elevados e dignos de maior atenção, além de não levar em conta que as populações anômala e de *background* têm, em muitos casos, superposição. A utilização intensiva desta regra pode ser resultado de um mau entendimento, pois o que se busca, na realidade, são *outliers* e não os extremos de uma distribuição estatística, seja ela normal ou lognormal (casos mais comuns em geoquímica). Se estivermos com uma distribuição normal e utilizarmos a abordagem média mais dois desvios padrão estamos assumindo que os 2,5% superiores são anômalos, o que se configura em uma tomada de decisão arbitrária baseada unicamente em um modelo estatístico. Apesar das limitações, entretanto, deve ser ressaltado que onde não houver evidência óbvia da presença de processos mineralizantes nos dados, e uma única população estiver sendo tratada, o uso de média mais dois desvios padrão como *threshold* pode ser um fator de segurança útil para isolar alguns dados para posterior avaliação.

DISTRIBUIÇÃO NORMAL

A curva normal é simétrica em relação à origem, sendo a área total sob a curva arbitrada como valendo 1 ou 100%, e a variável estudada tem uma probabilidade de ocorrência entre dois pontos igual à área sob a curva compreendida entre estes dois valores. Variáveis cujos resultados numéricos são resultado de diversos efeitos aleatórios pequenos e não relacionados têm distribuição aproximadamente normal. O ponto mais alto da curva concentra a média, a mediana e a moda da distribuição. A percentagem de valores em alguns intervalos são: 68,26% dos valores de uma variável aleatória normal estão dentro do intervalo compreendido entre as distâncias máximas de um desvio padrão positivo e um desvio padrão negativo a partir da sua média, 95,44% dos valores de uma variável aleatória normal estão dentro do intervalo compreendido entre as distâncias máximas de dois desvios padrão positivos ou negativos a partir da sua média, 99,72% dos valores de uma variável aleatória normal estão dentro do intervalo compreendido entre as distâncias máximas de três desvios padrão positivos ou negativos a partir da sua média.

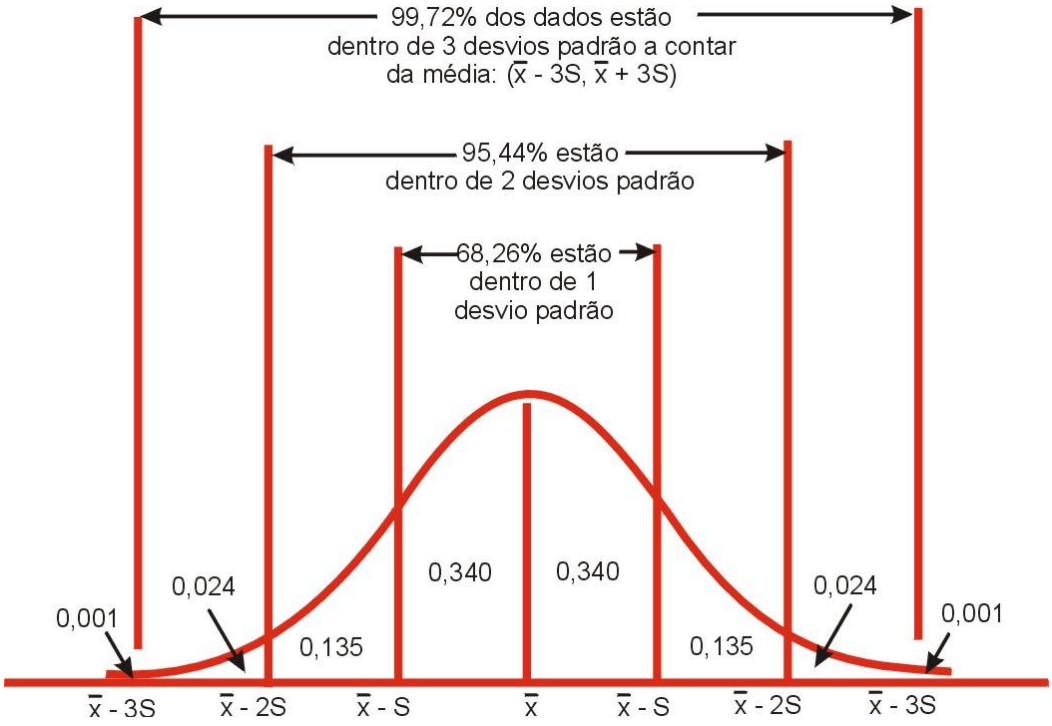


Figura 1 - Distribuição Normal

DISTRIBUIÇÃO Z

O escore padronizado ou escore **z** é o número de desvios padrão pelo qual um valor qualquer dista da média, e vale

$$z = (X_i - \bar{X}) / S$$

onde X_i representa cada observação, $\bar{X}$ representa a média e **S** o desvio padrão.

A estatística **z** tem distribuição normal **N(0,1)**, ou seja, com média igual a zero e variância igual a um. Quando se tem pequenas amostras – assim consideradas aquelas com menos de 30 observações – o **z** da distribuição normal não é adequado por não captar bem as diferenças entre os parâmetros populacionais e os amostrais, especialmente quando a dispersão populacional é grande. Em tais casos, ele é substituído pelo **t** da distribuição de Student.

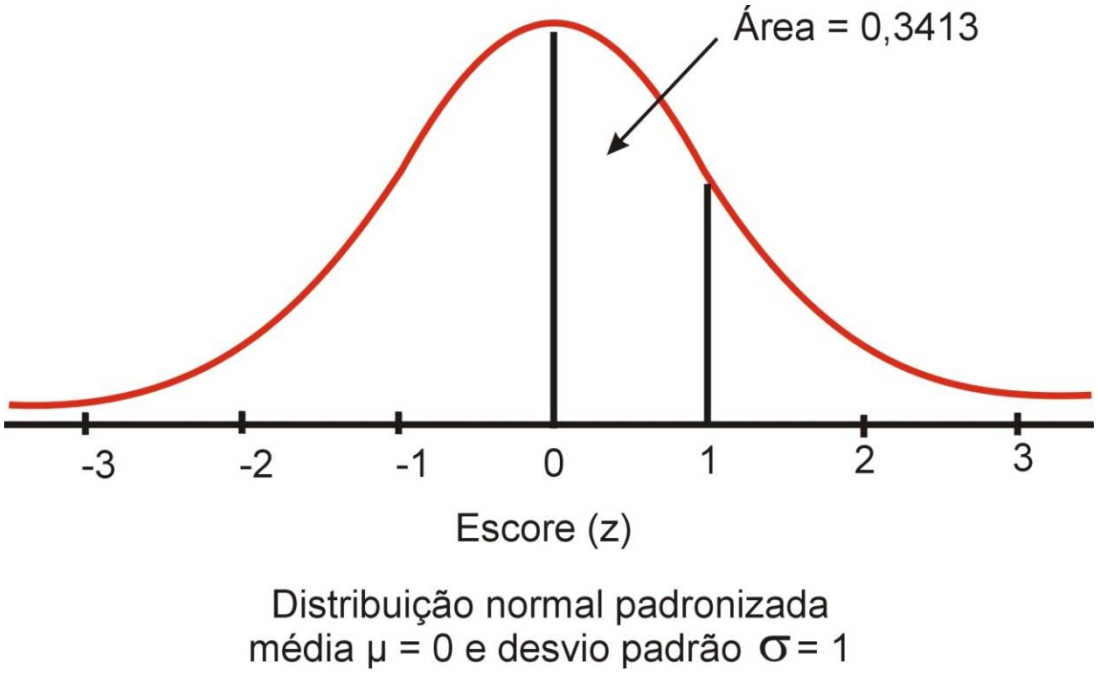


Figura 2 - Distribuição Normal Padronizada

NORMAL x LOGORMAL

Segundo Limpert *et al*. (2001) há algumas razões pelas quais as pessoas preferem a distribuição normal à lognormal, como a presença de simetria, que é um dos princípios básicos do modo de pensar do ser humano, a simplicidade, uma vez que a operação de adição é mais simples que a de multiplicação (no que diz respeito a esta razão, tome-se o caso de dois dados, e apenas como exemplo, já que a distribuição aqui exemplificada não se ajusta a nenhum dos dois modelos de distribuição: no caso aditivo eles variam de 2 a 12, e no multiplicativo variam de 1 a 36 com distribuição altamente assimétrica), e outra seria o fato de a distribuição normal ser conhecida e aplicada desde muito antes da lognormal. Distribuições assimétricas são particularmente comuns quando os valores médios são baixos e as variâncias são grandes, e muito comumente se ajustam bem às distribuições lognormais. Para dados normalmente distribuídos o intervalo ($\mu \pm \sigma$) cobre cerca de 68,3% da probabilidade e ($\mu \pm 2\sigma$) cobre cerca de 95,5%. Os valores correspondentes para quantidades lognormais são [μ^* / σ^*, $\mu^* . \sigma^*$] que contém 68,3% e [$\mu^* / (\sigma^*)^2$, $\mu^* . (\sigma^*)^2$], que contém 95,5% do total dos dados.

Para uma distribuição lognormal o mais preciso, isto é, assintoticamente mais eficiente método para estimar os parâmetros μ^* e σ^* repousa na logtransformação. A média e o desvio padrão empíricos dos logaritmos dos dados são calculados e então retrotransformados. O coeficiente de variação (CV) é um guia útil para a não normalidade, alternativamente a assimetria dos dados pode ser estimada, se o CV for maior que 100% gráficos em escala logarítmica devem ser preparados. Se o CV estiver entre 70% e 100% a

inspeção de gráficos em escala logarítmica será informativa. Outro guia útil é a razão valor máximo / valor mínimo, se exceder duas ordens de magnitude os gráficos logarítmicos serão informativos.

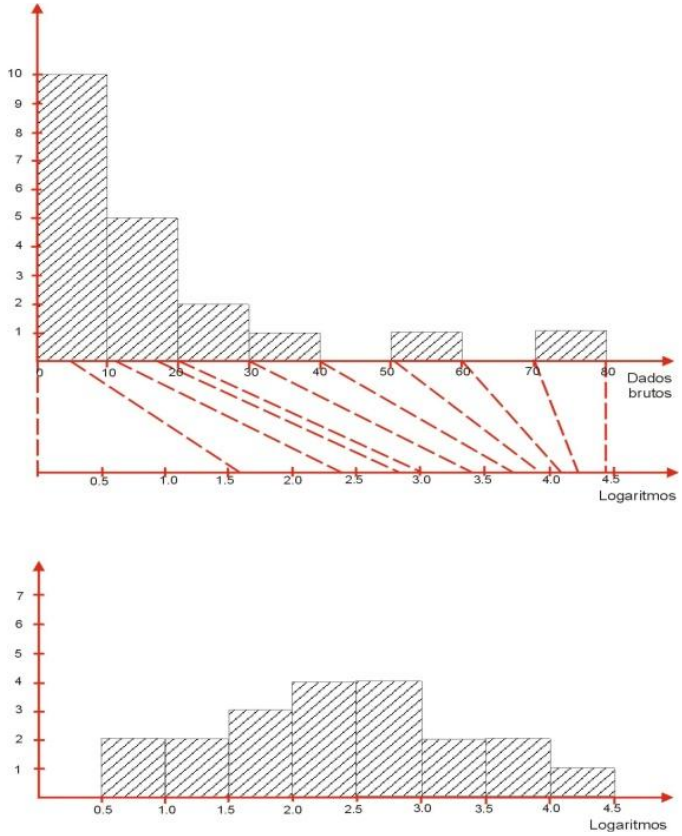


Figura 3 - Logtransformação

Mesmo após a conversão dos dados brutos para logaritmos alguns conjuntos de dados continuam apresentando forte assimetria, o que pode ser contornado testando-se se os dados se ajustam ou não ao modelo chamado de lognormal a três parâmetros, em que o terceiro parâmetro é uma constante **(C)** adicionada a cada um dos dados brutos, obtido da seguinte maneira:

$$C = \frac{\mathbf{Med} - \mathbf{p_1} \cdot \mathbf{p_2}}{\mathbf{p_1} + \mathbf{p_2} - \mathbf{2Med}}$$

em que **Med** representa a mediana dos dados brutos, $\mathbf{p_1}$ representa o valor correspondente a um percentil situado entre 5 e 20%, e $\mathbf{p_2}$ um valor correspondente a um percentil igual a $\mathbf{(100 - p_1)}$; o percentil $\mathbf{p_1}$ deve ser obtido por tentativas até que se obtenha o melhor ajuste da distribuição, sendo o percentil $\mathbf{p_2}$ obtido de modo automático pela diferença.

GRÁFICOS DE PROBABILIDADES

A análise gráfica da distribuição por meio da probabilidade acumulada (ou Q-Q *plot*) antes de definir os intervalos de *background* ou *threshold* é uma etapa muito importante, pois pode fornecer uma primeira visão, antes de cálculos numéricos, para as posições do *background* e do *threshold*. Nos gráficos Q-Q ou *Q-normal plots* (quantis da distribuição dos dados são plotados contra os quantis de uma distribuição hipotética como a normal) se pode observar desvios da normalidade e da lognormalidade, bem como a presença de múltiplas populações, assim como a presença de óbvios *outliers*.

Os gráficos de probabilidade são construídos com os valores acumulados de uma distribuição, com os valores da variável plotados contra os respectivos valores acumulados de freqüência. Este tipo de gráfico é muito sensitivo ao afastamento da normalidade e, em conseqüência, ao reconhecimento da presença de múltiplas populações no conjunto de dados. No caso de uma única população com distribuição normal os pontos do gráfico tendem a se dispor segundo uma reta; as “quebras” presentes nos gráficos de probabilidade não indicam, obrigatoriamente, a presença de múltiplas populações no nosso conjunto de dados, elas representam modificações nas características das freqüências acumuladas em diferentes intervalos, e isto deve ser estudado em detalhe. O que não se pode é “ver” coisas demais em um gráfico de probabilidade; bom alinhamento dos pontos não é uma garantia definitiva de uma boa estimativa, bem como a presença de “quebras” no gráfico não é um fator condenatório da estimativa. Nos gráficos acumulativos se representa os valores acumulados contra os valores superiores dos intervalos de classe.

ALGUMAS MEDIDAS DE VARIABILIDADE

Além das medidas de uso mais comum em Estatística, como variância e desvio padrão, outras medidas expressam a variabilidade de atributos, algumas delas comentadas a seguir.

O **coeficiente de variação (CV)** é o resultado da divisão do desvio padrão pela média aritmética, e seu valor mostra o quão maior (ou menor) o desvio padrão é da média. É uma grandeza útil para comparar distribuições de unidades diferentes. Em geral, quando **CV > 1** aconselha-se fazer transformação logarítmica dos dados. O coeficiente de variação dá uma idéia da regularidade ou homogeneidade das amostras que estão sendo estudadas. Valores elevados, na prática superiores a cerca de um, representam amostras com grande heterogeneidade, e valores abaixo de cerca de 0,4 refletem homogeneidade da amostra. Dentre estas últimas estariam espessuras de camadas sedimentares e teores de óxidos principais em rochas, além de vários casos de teores nos denominados depósitos minerais de alto teor. Á medida

que os teores diminuem ou a complexidade geológica do processo de formação da acumulação estudada aumenta o valor do **CV** (ou seja, a heterogeneidade) tende a aumentar, caso dos elementos metálicos e elementos raros na natureza.

A **assimetria** é o grau de desvio de uma curva no sentido horizontal, pode ser positivo, com excesso de valores altos, ou negativo, com predomínio de valores baixos.

A assimetria é dada pela fórmula

$$A_3 = \sum_{i=1}^{n} (X_i - \bar{X})^3 / n \cdot S^3$$

Para uma distribuição normal simétrica se tem $A_3 = 0$.

A assimetria é uma medida do grau de afastamento da média de uma distribuição em relação à sua moda e à sua mediana.

A **curtose** é o grau de achatamento de uma curva em relação a uma curva representativa de uma distribuição normal. Designa-se como leptocúrtica a curva com um pico elevado, platicúrtica a uma curva achatada e mesocúrtica a intermediária.

A curtose é dada por

$$A_4 = \sum_{i=1}^{n} (X_i - \bar{X})^4 / n \cdot S^4$$

Assimetria e curtose são medidas muito utilizadas em sedimentologia, mas também são guias úteis no momento de se definir pelo tipo de distribuição estatística dos dados de estudo.

O **erro padrão da média** dá uma idéia da precisão da estimativa da média, e pode ser obtido pela fórmula

$$S_X = S / \sqrt{n}$$

Assim, se uma amostra com 100 observações tiver um desvio padrão igual a 0,2 e média igual a 5, teremos erro padrão igual a 0,02 e dizemos que a estimativa para a média é $5 \pm 0{,}02$. Como se pode observar pela fórmula, a estimativa para a média se torna mais precisa (intervalo menor) com o aumento da quantidade de observações **(n).**

O **desvio entre quartis**, **IQR = (Q3 – Q1)**, é muitas vezes usado como medida de dispersão. Em distribuições simétricas a distância entre a mediana e o primeiro quartil é igual à distância entre a mediana e o terceiro quartil. A diferença entre o terceiro e o primeiro quartis é conhecida pelo nome de diferença interquartil (*Inter Quartile Range*, ou **IQR**, em inglês).

Uma medida de variabilidade de uso mais restrito é o **Desvio Absoluto da Média**, e vale

$$\sum \left| X - \overline{X} \right| / n$$

ou seja, se subtrai a média de cada dado, tomando seu valor absoluto, e se toma a média deste novo conjunto de valores. Como esta medida não eleva as parcelas ao quadrado as observações mais afastadas do valor médio afetam menos esta medida do que a variância e o desvio padrão.

Uma outra medida de variabilidade é o **Desvio Absoluto da Mediana** (MAD, de *Median Absolute Deviation*), e vale

$$\textbf{MAD = mediana } (\left| X - \textbf{mediana dos dados originais} \right|)$$

que é uma medida menos afetada por valores extremos, já que a mediana não sofre a mesma influência dos dados extremos que a média; por esta razão dados com valores extremos têm nesta medida uma estimativa mais estável de variabilidade do que medidas que utilizam média, variância e desvio padrão. Na determinação do MAD se utiliza o conceito de mediana duas vezes, uma diretamente sobre os dados originais e outra sobre as diferenças obtidas entre eles e a respectiva mediana, ou seja, sobre os resíduos obtidos a partir da mediana dos dados. O desvio padrão é uma medida útil para dados com distribuição normal, e desvios da normalidade o afastam da melhor condição de representação da variabilidade dos dados. A utilização da mediana e do MAD não apresenta a exigência de que os dados devam se ajustar a qualquer modelo de distribuição. Alguns autores propõem definir o valor do *threshold* a duas vezes o valor do MAD (*Median Absolute Deviation*) a partir da mediana. A abordagem MAD é melhor aplicada quando os dados contêm menos que 10% de *outliers*. Geralmente o método (mediana $\pm$ 2 MAD) dá o mais baixo *threshold*, identificando maior número de *outliers*, seguido pelo *boxplot*. O *threshold* definido pelo *boxplot* é em muitos casos próximo (mas menor) do que o obtido a partir de dados logtransformados com a utilização da regra (média $\pm$ 2 desvios padrão). Assim, pela utilização da regra (mediana $\pm$ 2 MAD) se obtém o *threshold* mais baixo, com o uso do *boxplot* ele fica mais elevado e pela aplicação da regra clássica (média $\pm$ desvios padrão) o *threshold* fica ainda maior.

Um exemplo comparando estes conceitos de variabilidade:

X	$\lvert X - \overline{X} \rvert$	$\lvert X - \text{mediana} \rvert$	$(X - \overline{X})^2$
1,2	0,5	0,3	0,25
1,4	0,3	0,1	0,09
1,5	0,2	0,0	0,04
1,6	0,1	0,1	0,01
2,8	1,1	1,3	1,21
Totais	2,2	1,8	1,60

Média = 1,7

Mediana = 1,5

Desvio Absoluto da Média = 0,44

Desvio Absoluto da Mediana = 0,36

Desvio Padrão = 0,57

Adicionamos a abordagem (mediana $\pm$ 2 MAD), o MAD é definido como o valor mediano dos desvios absolutos a partir da mediana de todos os dados, de acordo com Tukey (1977) porque, em adição ao *box-plot*, é a mais inteligível abordagem robusta, é uma analogia direta a (média $\pm$ 2 desvios padrão). O método clássico somente se mostra bom se não existirem *outliers*. É também usual a utilização do *boxplot* para a identificação de valores extremos. Uma característica importante da *boxplot* é o fato de que os limites para definição de *outliers* não são necessariamente simétricos ao redor do centro (mediana).

PRECISÃO, EXATIDÃO, REPETIBILIDADE, REPRODUTIBILIDADE

Levinson (1974) comenta que em exploração geoquímica a precisão, que ele define como a capacidade de reproduzir ou repetir o mesmo resultado, normalmente tem maior importância que a acuracidade, definida como sendo a aproximação a um determinado valor verdadeiro, pelo menos nas fases iniciais de um programa exploratório, e cita um exemplo em que valores de Zn se repetem em torno de 200 ppm, o que denota precisão, e que embora o valor verdadeiro seja 250 ppm não se gera um grande problema porque o que se tem em mãos é um conjunto de resultados medidos da mesma maneira, o que permite comparar resultados em diferentes locais de amostragem e, assim, delimitar anomalias, que são o interesse maior de uma fase inicial em exploração geoquímica.

A precisão mede o grau de concordância entre diversas medições feitas sobre o mesmo atributo, e serve para orientar os laboratórios sobre a dispersão do método adotado em certo procedimento analítico, e a exatidão ou acuracidade mede a concordância de vários resultados obtidos, por meio de seu valor médio, com o valor real, geralmente representado por um padrão de referência aceito como válido. As cartas de controle são ferramentas utilizadas para se aferir a precisão.

As figuras 4, 5 e 6 mostram, de forma esquemática, os conceitos de precisão e exatidão. Na figura 4 a situação **A** representa alta precisão e baixa acuracidade, ou seja, os valores se repetem mas não refletem o valor real do atributo estudado, a situação **B** representa imprecisão, uma vez que os valores não se repetem, mas acuracidade, uma vez que a média dos resultados obtidos reflete o valor real, e a situação **C** une altas precisão e acuracidade. A figura 5 é uma outra representação destes conceitos, e a figura 6 mostra, na situação **a**, precisão baixa associada a viés, ou erro sistemático, a situação **b** retrata baixa precisão mas sem viés, a situação **c** mostra precisão elevada com viés também elevado e a situação **d** representa precisão elevada e não enviezamento.

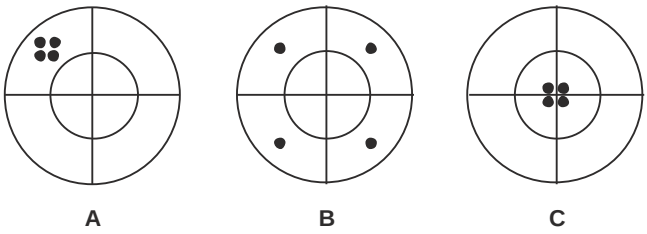


Figura 4 – Precisão e Exatidão

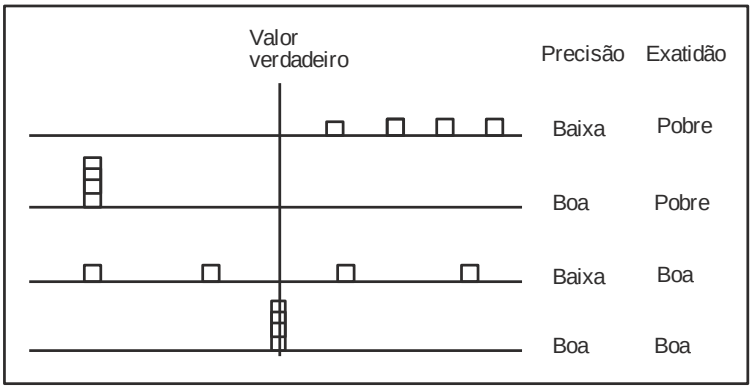


Figura 5 – Precisão e Exatidão

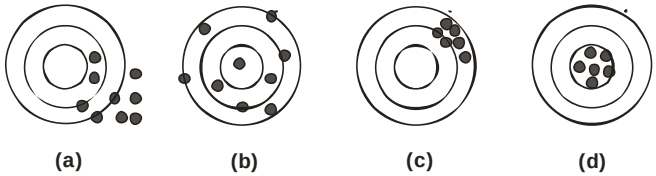


Figura 6 – Precisão e Exatidão

A **exatidão** pode ser expressa em termos percentuais, é o resultado da divisão da média da amostra de referência pela média verdadeira, ou seja, pelo valor verdadeiro da amostra de referência, 100% é o seu valor mais elevado.

A **precisão** se faz com amostras de controle, e pode ser expressa pelo coeficiente de variação, que é a divisão do desvio padrão pela média aritmética da amostra, e comparações entre precisões em diferentes níveis de concentração podem ser feitas pela aplicação do teste **F**, concluindo-se que se as diferenças neste teste não forem significativas ao nível de confiança escolhido a precisão do método se mantém para todo o intervalo de medição.

Repetibilidade é definida como sendo a diferença máxima aceitável entre medições feitas no mesmo dia, sobre o mesmo material, na bibliografia especializada é geralmente representada por **r**. A repetibilidade mede a concordância entre valores medidos com o mesmo método pela mesma pessoa, pelo mesmo equipamento, ou em mesma época. O valor de **r** é definido (para quantidades de dados iguais ou maiores que 10 e para 95% de confiabilidade) como sendo

$$\mathbf{r = 2{,}8 \,.\, Sr}$$

onde 2,8 é resultado da operação $2 \,.\, \sqrt{2}$, valor oriundo da distribuição normal, e **Sr** representa o desvio padrão dos resultados obtidos. Se trabalharmos com grau diferente de confiabilidade se usa

$$\mathbf{r = t \,.\, \sqrt{2} \,.\, Sr}$$

sendo o valor de **t** de Student relativo a um determinado α e um certo número de graus de liberdade.

Reprodutibilidade é definida como sendo a maior diferença aceitável entre medições feitas em dias diferentes, normalmente representada por **R**. É uma medida da concordância entre os resultados alcançados pela aplicação do mesmo método em amostras analisadas por operadores diferentes, ou laboratórios diferentes, ou mesmo equipamentos e épocas diferentes. O valor de **R** é definido (para quantidades de dados iguais ou maiores que 8 e para 95% de confiabilidade) como sendo

$$\mathbf{R = 2{,}8 \,.\, SR}$$

onde 2,8 é resultado da operação 2 . $\sqrt{2}$, valor oriundo da distribuição normal, e **SR** representa o desvio padrão dos resultados obtidos. Se trabalharmos com grau diferente de confiabilidade se usa

$$\mathbf{R = t \, . \, \sqrt{2} \, . \, SR}$$

sendo o valor de **t** de Student relativo a um determinado α e um certo número de graus de liberdade.

INTERVALOS DE CONFIANÇA DA MÉDIA

É um intervalo baseado em observações de uma amostra e construído de modo que haja uma probabilidade especificada de o verdadeiro valor desconhecido de um parâmetro estar contido neste intervalo; nível de confiança é a probabilidade de o intervalo conter o verdadeiro valor do parâmetro. É um intervalo real centrado na estimativa pontual que deverá conter o parâmetro com determinada probabilidade. A probabilidade de o intervalo conter o parâmetro estimado é denominado nível de confiança associado ao intervalo, cuja notação mais usual é **(1 - α)**. O que se afirma é que o percentual escolhido (68,26, 95,44 ou 99,72, por exemplo) representa a percentagem das vezes de que os intervalos contenham a verdadeira média, o que não é o mesmo que afirmar que esta é a probabilidade de ela cair dentro do intervalo, que é uma afirmação incorreta, pois ela é um número e, assim, está ou não está dentro do intervalo referido. Intervalo de 95% de confiança de que um valor desconhecido esteja entre 10 e 20, por exemplo, significa que obtivemos estes valores por um método que fornece resultados corretos em 95% das vezes.

Os limites dos intervalos são estabelecidos pelos valores

(estimativa – erro, estimativa + erro)

ou

$$\mathbf{[\ \bar{X} \pm (t \text{ ou } z) \, . \, S \, / \, \sqrt{n}\ ]}$$

onde $\bar{X}$ é a média da amostra, **S** o desvio padrão da mesma, **n** a quantidade de dados disponíveis e **t** ou **z** o processo utilizado, lembrando que estão relacionados a um valor de α pré-determinado.

TESTES DE HIPÓTESE

Ao se admitir uma hipótese qualquer como verdadeira, testa-se a sua validade para determinados níveis de confiança através dos chamados testes de hipótese. Ao se trabalhar sobre testes de hipótese dois tipos de erro podem estar presentes, e são denominados de tipos I e II.

Se rejeitamos **H_0 – a hipótese nula** - (e aceitamos **H_a, a hipótese alternativa**) quando, de fato, **H_0** é verdadeira, cometemos um erro tipo I. Erro do tipo I se chama nível de significância e se anota por α, que é a probabilidade de o teste rejeitar a hipótese nula **H_0** quando ela é, na verdade, verdadeira.

Se aceitamos **H_0** (rejeitamos **H_a**) quando, de fato, **H_a** é verdadeira, cometemos um erro tipo II. Erro do tipo II consiste em não rejeitar a hipótese nula quando ela for falsa, β é a probabilidade do erro do tipo II.

O ideal seria se estes dois valores pudessem valer zero, o que não é possível, o mais usual em um estudo é a fixação prévia do valor de α e **n** (tamanho da amostra), de modo que β fica determinado. Para um α fixo um aumento de **n** ocasiona uma redução de β, isto é, amostra maior reduz a chance de cometermos o erro de não rejeitar a hipótese nula quando ela for falsa, para um tamanho fixo uma diminuição de α acarreta um aumento de β, e reciprocamente um aumento de α acarreta uma diminuição de β; para reduzir ambos somente com aumento do tamanho da amostra. Eles são inversamente relacionados – à medida que um cresce o outro diminui. Os testes de hipótese servem para provar que uma hipótese é falsa, mas não ajudam muito a provar que seja verdadeira. Mesmo que aceitemos a hipótese nula isto não significa que não haja outra hipótese igualmente válida para os dados; para nos convencermos de que nossa hipótese seja realmente verdadeira devemos ter condições para rejeitar todas as hipóteses competidoras.

Se a diferença entre o que esperamos de acordo com a hipótese nula e o que observamos em uma amostra é demasiadamente grande para ser razoavelmente atribuída ao acaso, rejeitamos a hipótese nula. Se a diferença entre o que esperamos e o que observamos é suficientemente pequena para ser atribuída ao acaso dizemos que o resultado não é estatisticamente significativo, ou simplesmente não é significativo. Aceitamos, então, a hipótese nula. Chamamos hipótese nula a qualquer hipótese estabelecida especificamente para vermos se ela pode ser rejeitada. A razão da escolha da hipótese nula de forma que sua rejeição prove o que queremos afirmar, é que geralmente é muito mais fácil provar que algo é falso do que provar que algo é verdadeiro. Com relação à escolha entre **z** e **t**, tem-se que, para amostras com reduzido número de observações (30 ou menos) a aproximação à distribuição normal não é a mais adequada, e usamos, então, a distribuição **t**. A diferença principal entre elas é que a distribuição **t** tem área maior nas caudas, o que resultará que, a um nível de confiança estabelecido, o valor de **t** será um pouco superior ao valor respectivo de **z**. Por outro lado, a distribuição normal não depende do tamanho da amostra, e a distribuição **t** depende. Na escala **z** tomamos a média como origem e o desvio padrão como medida de afastamento a partir da média; assim, **z** assume valores negativos para valores

da variável inferiores à média e positivos para valores superiores à média; a média da distribuição vale zero, pois a média dista zero de si mesma.

Para pequenas amostras **(n < 30)** tanto **t** como **z** exigem que a população tenha distribuição normal, o que, para amostras maiores, não é imperativo. Se **n** for maior ou igual a 30 o valor de **t** pode ser aproximado pelo valor de **z**, pois que são aproximadamente iguais. A estatística **t** utiliza a média amostral, que se torna mais e mais próxima da normal na medida em que o tamanho da amostra aumenta, mesmo quando a população não tem distribuição normal.

O procedimento geral para se aplicar um teste de hipótese é o que segue (em que o valor observado $\mathbf{z_{calc}}$ ou $\mathbf{t_{calc}}$ será comparado com o valor crítico $\mathbf{z_{tab}}$ ou $\mathbf{t_{tab}}$ correspondente ao nível de significância α):

1) estabelecer $\mathbf{H_0}$ e $\mathbf{H_a}$ (hipóteses nula e alternativa, respectivamente)

2) estabelecer α relacionado a intervalo de confiança

3) dependendo de **n** escolher **z** ou **t**; a partir do α se chega ao $\mathbf{z_{tab}}$ ou $\mathbf{t_{tab}}$ utilizando a tabela **z** ou **t**

4) se calcula

$$\mathbf{S_a^2 = [(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2] / (n_1 + n_2 - 2)}$$

e a estatística do teste é dada por

$$\mathbf{z_{calc}\ ou\ t_{calc} = |\ (Média1 - Média2)\ |\ /\ S_a}\sqrt{[(1/n_1) + (1/n_2)]}$$

5) comparar $\mathbf{z_{calc}}$ ou $\mathbf{t_{calc}}$ com $\mathbf{z_{tab}}$ ou $\mathbf{t_{tab}}$ correspondente

6) se os valores calculados forem maiores que os tabelados se rejeita $\mathbf{H_0}$, caso contrário se aceita $\mathbf{H_0}$

TESTES BIVARIADOS E UNIVARIADOS

O teste mais comumente utilizado é o teste bivariado, que é o teste adotado quando não existe razão especial para esperar que as médias ou desvios padrão de duas amostras sejam diferentes. No caso de se suspeitar que a média ou o desvio padrão varie em apenas uma direção se deve usar o teste univariado. Neste caso dobra-se a probabilidade de acontecer o evento esperado. A incerteza de 5%, por exemplo, no teste bivariado, ou a probabilidade de 5% de que o valor crítico seja ultrapassado, será dividida em duas caudas, ou seja, 2,5% no final de cada cauda.

A tabela com apenas um lado da curva a um nível de confiança de 95% equivale a uma tabela da curva de dois lados a um nível de confiança

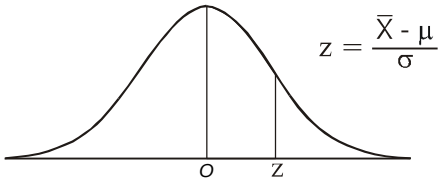
de 90%. Pode ocorrer que num teste o limite de 2,5% não tenha sido ultrapassado e então, na seqüência, esse limite de 2,5% foi ampliado para 5% e pode ser ultrapassado com maior facilidade.

De forma resumida se tem que o teste direcional ou **one-tailed** (unicaudal) é o que testa se a média amostral é diferente ou não da média da população em uma direção específica, menor ou maior, e o teste não direcional ou **two-tailed** (bicaudal) testa somente se um valor é diferente de outro, independentemente de ser maior ou menor. É importante a decisão sobre se o teste a ser aplicado será direcional ou não direcional, pois um valor que cai na região de rejeição em um teste unicaudal pode não cair na região de rejeição em um teste bicaudal. Na prática se usa o unicaudal quando se tem uma boa razão técnica para esperar que a diferença deva ser em uma direção particular. Ao se utilizar o teste bicaudal (para α = 0,05, por exemplo) é necessário procurar o valor tabelado sob a probabilidade de 0,025.

É importante salientar que só se pode estabelecer uma correspondência entre um intervalo de confiança e um teste de hipótese quando o teste for bilateral. Para mesmo α e mesmo **n** (quantidade de observações da amostra) a prova unicaudal é mais precisa do que a bicaudal, e o teste bicaudal só deve ser utilizado quando não houver informação sobre o sentido da diferença (se para mais ou para menos).

Figura 7 - Testes de Hipótese

Tabela 1 – Distribuição Normal Padronizada z

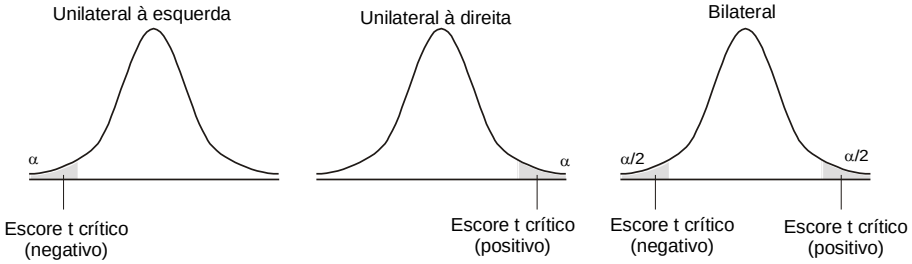


Distribuição Normal Padronizada (z)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3S88	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4S30	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4S68	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1 ou mais	0,4999									

Obs.: Cada valor da tabela representa a proporção da área total situada entre zero e o valor positivo de *z*, sendo as áreas relativas aos valores negativos de *z* obtidas por simetria. Se o valor de *z* for 1,96 a área entre zero e *z* valerá 0,4750, e a área total entre –*z* e *z* valerá 0,95. Para *z* = 2,57 se tem uma área total de aproximadamente 0,99 entre –*z* e *z*.

Tabela 2 – Distribuição t (de Student)



Distribuição t

α

Graus de liberdade	0,005 (unilateral) 0,01 (bilateral)	0,01 (unilateral) 0,02 (bilateral)	0,025 (unilateral) 0,05 (bilateral)	0,05 (unilateral) 0,10 (bilateral)	0,10 (unilateral) 0,20 (bilateral)	0,25 (unilateral) 0,50 (bilateral)
1	63,657	31,821	12,706	6,314	3,078	1,000
2	9,925	6,965	4,303	2,920	1,886	0,816
3	5,841	4,541	3,182	2,353	1,638	0,765
4	4,604	3,747	2,776	2,132	1,533	0,741
5	4,032	3,365	2,571	2,015	1,476	0,727
6	3,707	3,143	2,447	1,943	1,440	0,718
7	3,500	2,998	2,365	1,895	1,415	0,711
8	3,355	2,896	2,306	1,860	1,397	0,706
9	3,250	2,821	2,262	1,833	1,383	0,703
10	3,169	2,764	2,228	1,812	1,372	0,700
11	3,106	2,718	2,201	1,796	1,363	0,697
12	3,054	2,681	2,179	1,782	1,356	0,696
13	3,012	2,650	2,160	1,771	1,350	0,694
14	2,977	2,625	2,145	1,761	1,345	0,692
15	2,947	2,602	2,132	1,753	1,341	0,691
16	2,921	2,584	2,120	1,746	1,337	0,690
17	2,898	2,567	2,110	1,740	1,333	0,689
18	2,878	2,552	2,101	1,734	1,330	0,688
19	2,861	2,540	2,093	1,729	1,328	0,688
20	2,845	2,428	2,086	1,725	1,325	0,687
21	2,831	2,518	2,080	1,721	1,323	0,686
22	2,819	2,508	2,074	1,717	1,321	0,686
23	2,807	2,500	2,069	1,714	1,320	0,685
24	2,797	2,492	2,064	1,711	1,318	0,685
25	2,787	2,485	2,060	1,708	1,316	0,684
26	2,779	2,479	2,056	1,706	1,315	0,684
27	2,771	2,473	2,052	1,703	1,314	0,684
28	2,763	2,467	2,048	1,701	1,313	0,683
29	2,756	2,462	2,045	1,699	1,311	0,683
Grande (z)	2,575	2,327	1,960	1,645	1,282	0,675

Exemplo – Se o teste for unilateral, para a área marcada representar 0,05 da área total o valor de *t* será, com 20 graus de liberdade, igual a 1,725 (se o teste for bilateral o valor de *t* será 2,086)

DISTRIBUIÇÃO F

Para testar a hipótese de igual dispersão (medida pela variância da distribuição) se utiliza a razão das variâncias amostrais, o que pode ser feito pela estatística **F**. A distribuição **F** se constitui, na realidade, numa família de distribuições em que se exige o conhecimento de dois parâmetros, que são os graus de liberdade das variâncias amostrais no numerador e no denominador. O número de graus de liberdade do numerador é sempre mencionado em primeiro lugar, e a permuta dos números de graus de liberdade modifica a distribuição, daí ser importante manter a ordem, pois trocas alteram o valor obtido para a estatística **F**. As distribuições **F** assimétricas à direita, e como as variâncias amostrais não podem ser negativas a estatística **F** assume apenas valores positivos, a distribuição **F** não tem probabilidades correspondentes a valores negativos. O pico da curva de densidade de **F** está próximo de 1. Quando as duas populações têm o mesmo desvio padrão, as duas variâncias amostrais têm aproximadamente o mesmo valor, de modo que **F** toma um valor próximo de 1. Valores de **F** muito afastados de 1 em uma ou outra direção constituem evidência contra a hipótese de desvios padrão iguais.

Para executar o teste se determina, primeiramente, a estatística do teste, que vale

F = maior variância / menor variância

ou seja, definimos como população 1 a que apresenta a maior das variâncias observadas, o que implica que o valor obtido para **F** será sempre igual ou maior que um. No seguimento se compara o valor de **F** com o valor crítico de tabela, e se duplica os níveis de significância da tabela para obter o nível de significância para o teste **F** bicaudal; por fim se faz um teste de hipótese pela comparação entre os valores calculado e tabelado, testando-se se a diferença entre as variâncias comparadas é significativa ou não ao nível de significância escolhido. Para identificar uma distribuição **F** específica devemos indicar os graus de liberdade do numerador e do denominador. Para determinar o valor crítico escolhemos α para teste unilateral ou metade dele para bilateral, e buscamos a intersecção da coluna com os graus de liberdade da primeira amostra com a linha com os graus de liberdade da segunda.

Ao testarmos uma hipótese sobre as variâncias de duas populações admitimos que elas são independentes e distribuídas normalmente, o que é muito importante dado o fato de este teste ser extremamente sensível a desvios da normalidade, o que pode levar à rejeição de uma hipótese nula pela razão errada.

CORRELAÇÃO LINEAR

O conceito de correlação se refere a uma associação numérica entre duas variáveis, não implicando necessariamente uma relação de causa e efeito ou mesmo a existência de uma estrutura com interesses práticos. Se a representação gráfica das duas variáveis em um sistema cartesiano resulta em pontos alinhados, se ajustando a uma reta, se está na presença de uma relação linear.

O coeficiente de correlação linear é representado por **r**, e os valores de **r** variam entre -1 (correlação inversa) e +1 (correlação direta), sendo o valor zero representativo de ausência de correlação linear. O coeficiente de correlação linear é uma medida da intensidade da relação linear entre duas variáveis, e mede o grau de relacionamento linear entre os dados emparelhados das variáveis X e Y em uma amostra, e recebe o nome de Momento-Produto de Pearson.

O valor $\mathbf{r^2}$ representa a parte da variância total de X e Y que pode ser explicada pela sua relação linear, e se tem

$$\mathbf{r^2 = variância\ explicada\ /\ variação\ total}$$

ou

$$\mathbf{r^2 = (variação\ total - variação\ não\ explicada)\ /\ (variação\ total)}$$

Assim, $\mathbf{r^2}$ é a proporção da variação total em Y explicada pelo ajuste da regressão, e é chamado coeficiente de determinação. Se verificarmos, entre duas variáveis X e Y, um valor de **r** igual a 0,7 ter-se-á $\mathbf{r^2}$ = 0,49, ou seja, o grau de dependência de Y em relação a X será de 49%; isto significa que 51% da variação total permanece não explicada.

Muitas variáveis são expressas em percentagens e somam 100%, o que dificulta estudos de correlação linear. Quando um cresce o outro obrigatoriamente decresce, o que resulta, como conseqüência, altos valores negativos para **r**. Igualmente difíceis de interpretar são as razões. Sempre que possível deve-se evitar o estudo de correlações entre razões diretamente. O fato de termos um valor de **r** igual a zero não significa que não exista correlação entre as variáveis; significa, isto sim, que elas não têm, entre si, apenas a correlação linear.

REGRESSÃO LINEAR

Só se deve utilizar a regressão se a correlação entre as variáveis for significativa. A correlação mede a força ou grau de relacionamento entre duas variáveis, e a regressão fornece uma equação que descreve o relacionamento entre elas em termos matemáticos. O ideal seria a previsão, em função da

relação existente, dos valores exatos de uma variável, mas o que se consegue é apenas prever valores médios, ou valores esperados.

O método dos mínimos quadrados é um método de ajuste de pontos a uma reta, e se baseia em que a reta resultante do ajuste seja tal que a soma dos quadrados das distâncias verticais dos pontos à reta seja mínima, reta esta que recebe o nome de reta dos mínimos quadrados, reta de regressão ou reta de regressão estimada, sendo os valores de **a** e **b** da equação da reta estimados com base em dados amostrais.

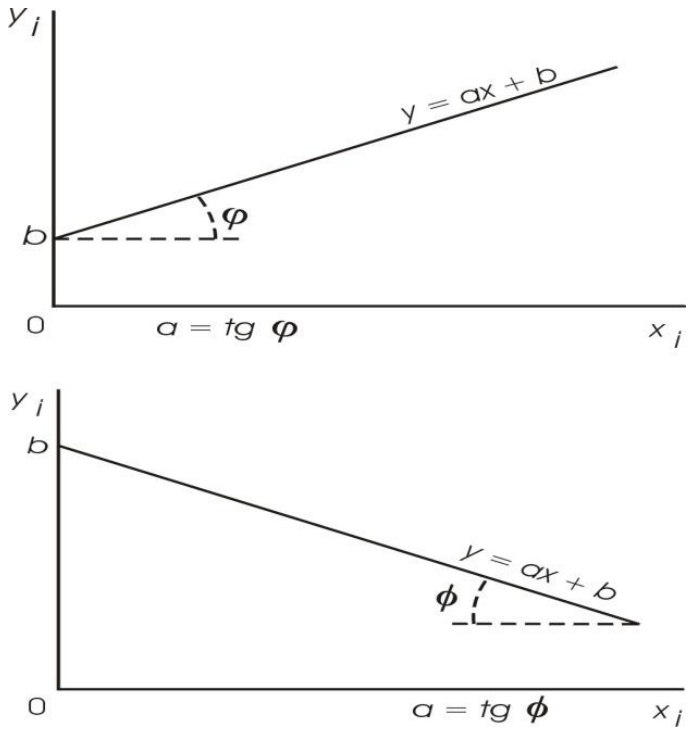


Figura 8 - Reta de Regressão

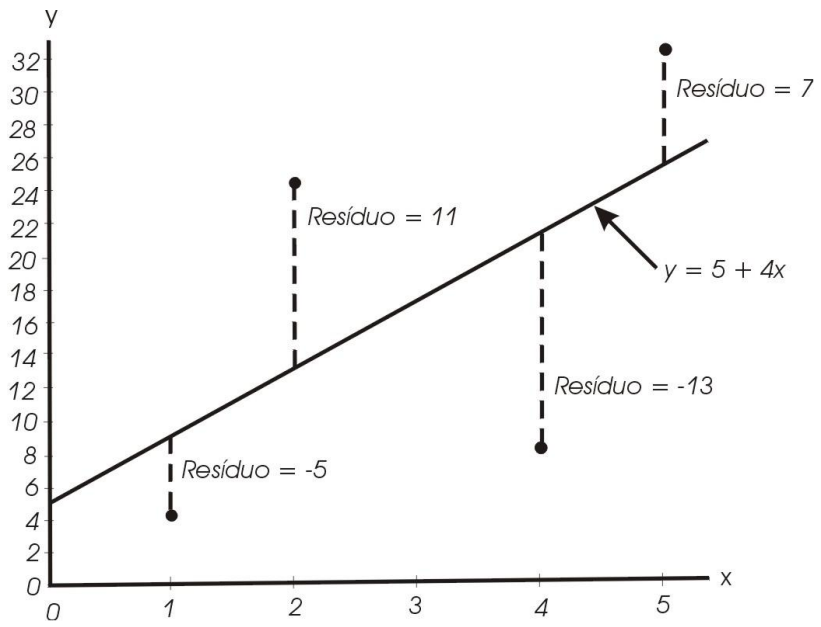


Figura 9 - Mínimos Quadrados e Resíduos

OUTLIERS

Em geoquímica o interesse maior ocorre na determinação de *outliers* como indicadores de processos geoquímicos raros (como mineralizações, por exemplo). *Outliers* não influenciam métodos robustos, e métodos não paramétricos não se baseiam em suposições de distribuição, por isso são preferíveis aos métodos clássicos. O fato de um determinado elemento apresentar muitos *outliers* pode, por si só, indicar a existência de mineralizações. De um modo geral se elimina os *outliers* quando eles representam erros óbvios, mas frequentemente eles representam anomalias interessantes que merecem estudo mais detalhado, na verdade para alguns conjuntos de dados os *outliers* são a característica mais importante. A identificação de valores pertencentes a um conjunto de dados que possam ser caracterizados como *outliers* (ou valores aberrantes, segundo alguns autores de língua portuguesa), bem como o tratamento que se deve dar a eles é tema importante no tratamento estatístico de dados. *Outliers* são tão diferentes dos demais valores disponíveis para estudo que se pode suspeitar da presença de uma observação que não pertence ao grupo de estudo.

Existem vários métodos para se fazer este estudo, um deles propõe classificar como *outliers* os dados que estão na área correspondente a 1% da parte superior da curva de distribuição, o que pressupõe que sempre se terá *outliers* entre os dados, outro propõe eliminar todos os valores que estejam a uma distância da média maior que três desvios padrão (muito similar ao anterior), o que tem limitações porque tanto a média quanto o desvio padrão são afetados pelos *outliers*, outro propõe transformar os dados (logaritmos, raiz quadrada). Um ponto a se levar em conta é que ao se adotar uma

transformação a relação da nova variável com as demais variáveis originais sofre modificações. Outra opção é a utilização de métodos não paramétricos, com a utilização de estatística robusta (mediana, *inter quartile range*). Neste caso, chamando-se **Q3** ao terceiro quartil, **Q1** ao primeiro quartil, e **IQR** à diferença **(Q3 – Q1)**, sendo classificados como *outliers* aos valores inferiores a **(Q1 – 1,5 IQR)** e superiores a **(Q3 + 1,5 IQR)**, técnica esta que será tratada com mais detalhe adiante.

Um fato sem contestação é que a presença de *outliers* nos dados afeta as conclusões tiradas a partir do exame estatístico, são afetados os valores médios, as medidas de dispersão e as correlações com outras variáveis de interesse, afetando inclusive estudos multivariados em que a variável em questão estiver incluída. O que não se deve fazer é simplesmente ignorar a presença dos *outliers*, nem tampouco apenas eliminá-los sem descobrir a razão de sua presença nos dados, sob pena de chegarmos a conclusões que não dizem respeito nem ao grosso dos dados nem aos valores mais elevados. Até alguns anos atrás as medições que se referiam à existência do buraco na camada de ozônio eram tratadas como *outliers* e eram automaticamente descartadas. Dentre as várias possíveis fontes deste tipo de dado se pode citar a presença de erros analíticos, contaminação, erros de digitação e/ou transcrição de resultados, e erros de interpretação, como classificar erroneamente determinado grupo, incluindo seus valores em outro grupo.

Apresentamos de forma resumida, a seguir, diversas técnicas ou formas de classificar um determinado valor como *outlier* ou de determinação de limiares.

TESTE DO ESCORE z MODIFICADO

Este teste tem sido usado de forma mais extensiva que o teste que considera como *outlier* simplesmente os valores que superam a soma da média aritmética com três desvios padrão, ou a média menos três desvios padrão, pois que tanto a média como o desvio padrão são, já, afetados pela presença do *outlier*. O teste do escore **z** modificado usa estimadores robustos, como a mediana, o que garante não terem sido os valores utilizados para definir um *outlier* afetados pelo mesmo. Por meio de um exemplo (tabela que segue) montaremos uma verificação da presença de um *outlier* com este teste.

Dado original (Xi)	$\lvert X_i - Xm \rvert$	z_i
3,2	0,1	- 0,34
3,3	0,0	0,00
8,1	4,8	16,19
3,2	0,1	- 0,34
2,9	0,4	- 1,35
3,7	0,4	1,35
3,1	0,2	- 0,67
3,5	0,2	0,67
3,3	0,0	0,00
9,2	5,9	19,90

1 º passo - se calcula a mediana dos dados brutos, que vale 3,3

2º passo – se determina a coluna com os valores dos desvios absolutos, definida por $\lvert \mathbf{X_i - X_m} \rvert$

3º passo – se determina a média aritmética dos desvios absolutos (MAD), valores que constam da coluna criada no passo anterior, que vale 0,2 neste caso

4 º passo – se calcula os valores de **z** modificado para cada observação, gerando a coluna três da tabela anterior; este valor é representado por $\mathbf{z^*_i}$, que **vale** $\mathbf{z^*i = 0,6745 (X_i - X_m) / MAD}$

para a terceira observação se tem $\mathbf{z^*_i}$ = 0,6745 (8,1 – 3,3) / 0,2 = 16,19

para a quarta observação se tem $\mathbf{z^*_i}$ = 0,6745 (3,2 – 3,3) / 0,2 = - 0,34

para a décima observação se tem $\mathbf{z^*_i}$ = 0,6745 (9,2 – 3,3) / 0,2 = 19,90

5º passo – se considera como *outliers* valores de $|\mathbf{z^*_i}| > 3,5$, ou seja, no caso estudado são considerados *outliers* os valores relativos a 16,19 e 19,90 da terceira coluna.

TESTE DE GRUBBS

Este teste é utilizado para dados que seguem a distribuição lognormal. Usaremos um exemplo para a sua utilização. É um teste definido como sendo útil principalmente para testar variabilidade entre laboratórios.

Dado original (X_i)	Ln (X_i)	Com *rank*
2,15	0,77	0,77
11,76	2,46	1,14
5,08	1,63	1,63
3,12	1,14	2,19
12,87	2,55	2,46
32,13	3,47	2,55
219	5,39	2,98
19,69	2,98	3,47
179	5,19	3,87
9609	9,17	4,31
327	5,79	4,62
74,2	4,31	5,19
102	4,62	5,39
47,8	3,87	5,79
8,97	2,19	9,17

1º passo – calcular a média e o desvio padrão dos dados já transformados em logaritmos naturais, respectivamente 3,70 e 2,17

2º passo – colocar os dados logtransformados em ordem crescente, com *rank*

3º passo – se houver suspeita de *outlier* para o menor valor se faz

$$\mathbf{\tau = [Média - X_1] / S}$$

se houver suspeita de *outlier* para o maior valor se faz

$$\tau = [X_n - \text{Média}] / S$$

No presente exemplo, suspeitando-se do maior valor se tem

$$\tau_{15} = [9,17 - 3,70] / 2,17 = 2,52$$

4º passo – para um α = 0,05 se determina o τ crítico para **n** = 15, no caso 2,409

5º passo – se o valor calculado for maior que o crítico se rejeita a hipótese nula e se conclui que o dado testado é um *outlier*; no caso presente, se rejeita a hipótese nula e o valor testado é um *outlier*. A tabela 3 contém os valores críticos para o teste de Grubbs, com α valendo 0,10, 0,05, 0,025, 0,01 e 0,005, unicaudais, ao se usar teste bicaudal se deve adotar a mesma tabela com o dobro das probabilidades α; esta tabela tem incrementos unitários para tamanhos de amostra entre 3 e 40 observações e incrementos de 10 unidades entre amostras com 40 a 140 observações.

n \ α	0.10	0.05	0.025	0.01	0.005
3	1.148	1.153	1.155	1.155	1.155
4	1.425	1.463	1.481	1.492	1.496
5	1.602	1.672	1.715	1.749	1.764
6	1.729	1.822	1.887	1.944	1.973
7	1.828	1.938	2.020	2.097	2.139
8	1.909	2.032	2.126	2.221	2.274
9	1.977	2.110	2.215	2.323	2.387
10	2.036	2.176	2.290	2.410	2.482
11	2.088	2.234	2.355	2.485	2.564
12	2.134	2.285	2.412	2.550	2.636
13	2.175	2.331	2.462	2.607	2.699
14	2.213	2.371	2.507	2.659	2.755
15	2.247	2.409	2.549	2.705	2.806
16	2.279	2.443	2.585	2.747	2.852
17	2.309	2.475	2.620	2.785	2.894
18	2.335	2.504	2.651	2.821	2.932
19	2.361	2.532	2.681	2.854	2.968
20	2.385	2.557	2.709	2.884	3.001
21	2.408	2.580	2.733	2.912	3.031
22	2.429	2.603	2.758	2.939	3.060
23	2.448	2.624	2.781	2.963	3.087
24	2.467	2.644	2.802	2.987	3.112
25	2.486	2.663	2.822	3.009	3.135
26	2.502	2.681	2.841	3.029	3.157
27	2.519	2.698	2.859	3.049	3.178
28	2.534	2.714	2.876	3.068	3.199
29	2.549	2.730	2.893	3.085	3.218
30	2.563	2.745	2.908	3.103	3.236
31	2.577	2.759	2.924	3.119	3.253
32	2.591	2.773	2.938	3.135	3.270
33	2.604	2.786	2.952	3.150	3.286
34	2.616	2.799	2.965	3.164	3.301
35	2.628	2.811	2.979	3.178	3.316
36	2.639	2.823	2.991	3.191	3.330
37	2.650	2.835	3.003	3.204	3.343
38	2.661	2.846	3.014	3.216	3.356
39	2.671	2.857	3.025	3.228	3.369
40	2.682	2.866	3.036	3.240	3.381
50	2.768	2.956	3.128	3.336	3.483
60	2.837	3.025	3.199	3.411	3.560
70	2.893	3.082	3.257	3.471	3.622
80	2.940	3.130	3.305	3.521	3.673
90	2.981	3.171	3.347	3.563	3.716
100	3.017	3.207	3.383	3.600	3.754
110	3.049	3.239	3.415	3.632	3.787
120	3.078	3.267	3.444	3.662	3.817
130	3.104	3.294	3.470	3.688	3.843
140	3.129	3.318	3.493	3.712	3.867

Tabela 3 – Valores Críticos de Grubbs

TESTE DE DIXON

O teste de Dixon para valores extremos atenta para a diferença entre os valores máximo e mínimo e seus valores vizinhos, é gerada uma razão **r** à qual é atribuída uma certa distribuição. O teste de Dixon é usado mais comumente na detecção de pequenas quantidades de *outliers*, é recomendado quando o número de observações está entre 3 e 25; os dados são ordenados de modo crescente e uma estatística é computada para o maior ou menor valor, suspeito de ser um *outlier*. Depois de estabelecido um nível de significância se o compara com um valor de tabela, se for menor que certo valor crítico a hipótese nula não é rejeitada, ou seja, aceita-se a hipótese de não existência de *outliers*, se a hipótese nula for rejeitada (valor calculado maior que o valor crítico) se conclui que o valor testado é um *outlier*. Para testar a existência de outros *outliers* se repete o teste, mas o poder deste teste diminui à medida que o número de repetições do mesmo aumenta. Alguns autores citam que o teste de Dixon não é mais recomendado por haver melhores opções disponíveis. A aplicação do teste de Dixon está exemplificada a seguir.

1º passo – os dados devem ser ordenados de forma crescente, sendo o menor valor o de ordem 1, e o maior valor o de ordem **N**

2º passo – chama-se **Z** ao valor numérico do dado de ordem **N**, ou seja, **Z(1)** é o valor numérico do menor resultado e **Z(N)** é o valor numérico do resultado de maior valor numérico, **Z(N – 1)** é o valor do penúltimo dado em ordem crescente de valor numérico; ao se proceder ao teste (**Q**) se chama **QM** ao valor mais elevado (suspeito de ser *outlier*) e **Qm** ao valor menor (suspeito de ser *outlier*).

3º passo – procede-se ao teste de Dixon, de acordo com três situações:

havendo entre 3 e 7 observações

QM = [Z(N) – Z(N – 1)] / [Z(N) – Z(1)]

Qm = [Z(2) – Z(1)] / [Z(N) – Z(1)]

havendo entre 8 e 12 observações

QM = [Z(N) – Z(N – 1)] / [Z(N) – Z(2)]

Qm = [Z(2) – Z(1)] / [Z(N - 1) – Z(1)]

havendo entre 13 e 25 observações

QM = [Z(N) – Z(N – 2)] / [Z(N) – Z(3)]

Qm = [Z(3) – Z(1)] / [Z(N – 2) – Z(1)]

Havendo mais de 25 observações o teste não está definido, deve-se buscar outra solução. A tabela 4 apresenta os valores críticos do teste de Dixon para valores de α iguais a 0,10, 0,05 e 0,01 unicaudais, para o caso bicaudal deve-se usar os mesmos valores críticos mas duplicando as probabilidades nos cabeçalhos das colunas. Esta tabela é válida ao se aplicar o teste de Dixon para conjuntos de dados que se ajustem à distribuição normal.

n	α = 0,10	α = 0,05	α = 0,01
3	0,886	0,941	0,988
4	0,679	0,765	0,889
5	0,557	0,642	0,780
6	0,482	0,560	0,698
7	0,434	0,507	0,637
8	0,479	0,554	0,683
9	0,441	0,512	0,635
10	0,409	0,477	0,597
11	0,517	0,576	0,679
12	0,490	0,546	0,642
13	0,467	0,521	0,615
14	0,492	0,546	0,641
15	0,472	0,525	0,616
16	0,454	0,507	0,595
17	0,438	0,490	0,577
18	0,424	0,475	0,561
19	0,412	0,462	0,547
20	0,401	0,450	0,535
21	0,391	0,440	0,524
22	0,382	0,430	0,514
23	0,374	0,421	0,505
24	0,367	0,413	0,497
25	0,360	0,406	0,489

Tabela 4 - Valores Críticos de Dixon

TESTE DE COCHRAN

O teste de Cochran é definido como sendo um teste para estudar variabilidade interna de um laboratório. O teste de Cochran é definido pela estatística **C**, que vale

$$C = S^2máx \,/ \sum_{i=1}^{p} S_i^2$$

onde **Smáx** é o desvio padrão máximo no conjunto; a hipótese nula parte do princípio que a estatística **C** tem uma distribuição aproximada à de qui quadrado com **(m – 1)** graus de liberdade, onde ***m*** representa o número de variáveis. O teste de Cochran é afetado pela não normalidade dos dados, e usa uma tabela específica, a tabela de Cochran. O teste de Cochran é uma variante do teste **t** (de Student, que compara conjuntos cujas variabilidades não sejam muito diferentes entre si), quando as amostras apresentam diferenças de variabilidade, verificada por um teste **F**. Resumidamente o teste de Cochran exige a ordenação crescente para cada conjunto de duas repetições, aplicar a fórmula anteriormente apresentada e comparar o valor obtido com o valor tabelado para este teste, se o valor da fórmula for menor que o tabelado não há dispersão, se for maior que o tabelado se diz haver dispersão quanto à amplitude. Consideraremos os dados que seguem, tomados em amostras de oito observações cada, para exemplificar o teste de Cochran.

	A	B	C	D
Média	2,75	3,50	6,25	9,00
Variância	2,214	0,857	1,071	1,714
Desvio Padrão	1,488	0,926	1,035	1,309

O valor de **C** será 2,214 / 5,856 = 0,378 com 4 grupos e 7 graus de liberdade, que são respectivamente **k** e **(n – 1)**, quatro conjuntos e cada conjunto com oito valores. O valor crítico para **C**, considerado um α igual a 0,05, é de 0,5365. Em conclusão, se rejeita a hipótese nula de que as variâncias sejam iguais. Os valores críticos de Cochran estão na tabela 5.

$\alpha = 0,05$

Nº grupos	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
GL																	
1	9985	9669	9065	8412	7808	7271	6798	6385	6020	5410	4709	3894	3434	2929	2370	1737	0998
2	9750	8709	7679	6838	6161	5612	5157	4775	4450	3924	3346	2705	2354	1980	1567	1131	0632
3	9392	7977	6841	5981	5321	4800	4377	4027	3733	3264	2758	2205	1907	1593	1259	0895	0495
4	9057	7457	6287	5441	4803	4307	3910	3584	3311	2880	2419	1921	1656	1377	1082	0765	0419
5	8772	7071	5895	5065	4447	3974	3595	3286	3029	2624	2195	1735	1493	1237	0968	0682	0371
6	8534	6771	5598	4783	4184	3726	3362	3067	2823	2439	2034	1602	1374	1137	0887	0623	0337
7	8332	6530	5365	4564	3980	3535	3185	2901	2666	2299	1911	1501	1286	1061	0827	0583	0312
8	8159	6333	5175	4387	3817	3384	3043	2768	2541	2187	1815	1422	1216	1002	0780	0552	0292
9	8010	6167	5017	4241	3682	3259	2926	2659	2439	2098	1736	1357	1160	0958	0745	0520	0279
10	7880	6025	4884	4118	3568	3154	2829	2568	2353	2020	1671	1303	1113	0921	0713	0497	0266
16	7341	5466	4366	3645	3135	2756	2462	2226	2032	1737	1429	1108	0942	0771	0595	0411	0218
36	6602	4748	3720	3066	2612	2278	2022	1820	1655	1403	1144	0879	0743	0604	0462	0316	0165
144	5813	4031	3093	2513	2119	1833	1616	1446	1308	1100	0889	0675	0567	0457	0347	0234	0120
∞	5000	3333	2500	2000	1667	1429	1250	1111	1000	0833	0667	0500	0417	0333	0250	0167	0083

$\alpha = 0,01$

Nº grupos	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
GL																	
1	9999	9933	9676	9279	8828	8376	7945	7544	7175	6528	5747	4799	4247	3632	2940	2151	1225
2	9950	9423	8643	7885	7218	6644	6152	5727	5358	4751	4069	3297	2821	2412	1915	1371	0759
3	9794	8831	7814	6957	6258	5685	5209	4810	4469	3919	3317	2654	2295	1913	1508	1069	0585
4	9586	8335	7212	6329	5635	5080	4627	4251	3934	3428	2882	2288	1970	1635	1281	0902	0489
5	9373	7933	6761	5875	5195	4659	4226	3870	3572	3099	2593	2048	1759	1454	1135	0796	0429
6	9172	7606	6410	5531	4866	4347	3932	3592	3308	2861	2386	1877	1608	1327	1033	0722	0387
7	8988	7335	6129	5259	4608	4105	3704	3378	3106	2680	2228	1748	1495	1232	0957	0668	0357
8	8823	7107	5897	5037	4401	3911	3522	3207	2945	2535	2104	1646	1406	1157	0898	0625	0334
9	8674	6912	5702	4854	4229	3751	3373	3067	2813	2419	2002	1567	1388	1100	0853	0594	0316
10	8539	6743	5536	4697	4084	3616	3248	2950	2704	2320	1918	1501	1283	1054	0816	0567	0302
16	7949	6059	4884	4094	3529	3105	2779	2514	2297	1961	1612	1248	1060	0867	0668	0461	0242
36	7067	5153	4057	3351	2858	2494	2214	1992	1811	1535	1251	0960	0810	0658	0503	0344	0178
144	6062	4230	3251	2644	2229	1929	1700	1521	1376	1157	0934	0709	0595	0480	0363	0245	0125
∞	5000	3333	2500	2000	1667	1429	1250	1111	1000	0833	0667	0500	0417	0333	0250	0167	0083

Tabela 5 – Valores Críticos de Cochran

Estas tabelas contêm os **valores críticos (C)** do teste de Cochran para homogeneidade de variâncias de amostras de igual tamanho. Todos os valores das duas tabelas devem ser divididos por 10.000, ou seja, elas contêm apenas a parte decimal, a parte inteira vale sempre zero. Assim, na tabela relativa a $\alpha = 0,05$, para **graus de liberdade (GL)** valendo 5 e dois grupos o valor tabelado vale 0,8772.

Estas tabelas contêm os valores críticos (**C**) do teste de Cochran para homogeneidade de variâncias de amostras de igual tamanho. Todos os valores das duas tabelas devem ser divididos por 10.000, ou seja, elas contêm apenas a parte decimal, a parte inteira vale sempre zero. Assim, na tabela relativa a α = 0,05, para graus de liberdade (**GL**) valendo 5 e dois grupos o valor tabelado vale 0,8772.

TESTE DE DOERFFEL

É um teste de simples aplicação introduzido por Doerffel em 1967 e confirmado por Dean & Dixon em 1981, e citados por Wellmer (1998). É utilizado para pequenos conjuntos de dados, e é representado basicamente por

$$\mathbf{Q = (X_a - X_r) / R}$$

em que $\mathbf{X_a}$ é o valor que se suspeita seja um *outlier*, $\mathbf{X_r}$ é o valor adjacente (mais próximo) dele, **R** representa a amplitude dos dados (valor máximo – valor mínimo), e **Q** é o valor do teste. O valor testado será aceito se o **Q** calculado for inferior ao valor tabelado por Doerffel (1967) e por Dean & Dixon (1981), em tabela reproduzida a seguir.

n (tamanho)	Q de Doerffel (α = 0,05)	Q de Dean & Dixon (α = 0,05)
3	0,97	0,94
4	0,84	0,76
5	0,73	0,64
6	0,64	0,56
7	0,59	0,51
8	0,54	0,47
9	0,51	0,44
10	0,49	0,41

Doerffel havia, anteriormente (em 1962), proposto um método para a detecção de *outliers* baseado em um diagrama, o valor é considerado aberrante por este método se superar a soma (média aritmética + desvio padrão x ***g***), sendo que a média aritmética e o desvio padrão devem ser calculados sem o valor suspeito (pois que ele afeta estes valores), e o valor de ***g*** pode ser obtido a partir de um diagrama específico para isto, este valor representa o *threshold* de um valor aberrante; este diagrama está representado na figura 10.

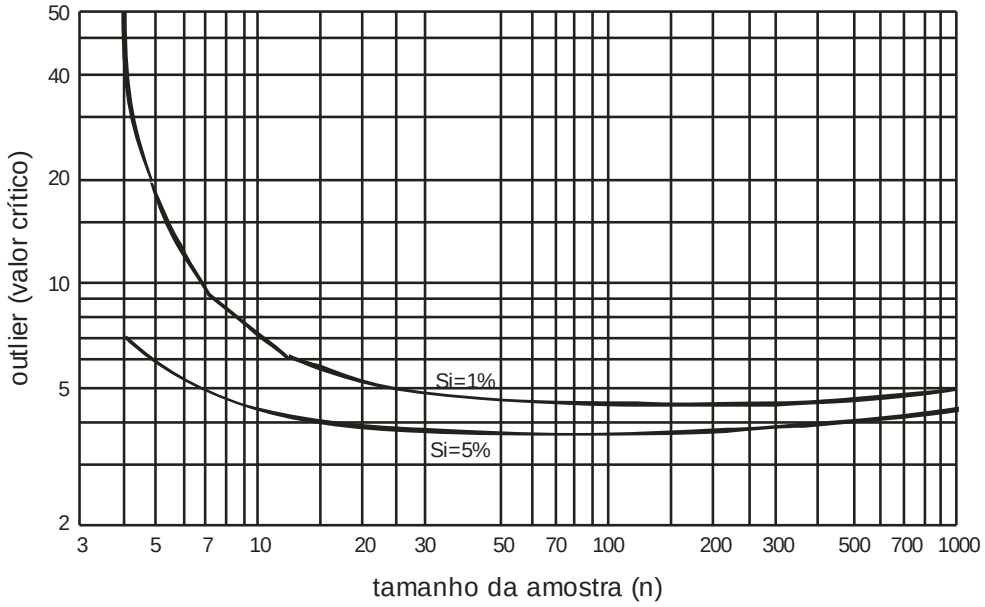


Figura 10 - Valores Críticos de Doerffel

Wellmer (1998) apresenta um exemplo de aplicação com os valores expressos em WO(%) iguais a 0,8, 1,4, 0,7, 2,4, 4,6, 2,1 e 1,5, sendo o valor 4,6% suspeito de ser um *outlier*, o valor adjacente a ele vale 2,4%. A aplicação do teste:

$$\mathbf{Q = (X_a - X_r) / R}$$

$$(4,6 - 2,4) / (4,6 - 0,7) = 2,2 / 3,9 = 0,56$$

Escolhendo-se o nível de significância ($\mathbf{S_i}$ = 5%) se verifica que o valor de **Q** é menor que o valor de $\mathbf{S_i}$ correspondente, logo o valor 4,6 é aceitável, ou seja, não deve ser classificado como um *outlier*.

GRÁFICOS DE PROBABILIDADES DE SINCLAIR

Os gráficos de probabilidade são muito sensitivos ao afastamento da normalidade e ao reconhecimento de múltiplas populações. O exame de muitos destes gráficos mostra que um ponto de inflexão ocorre na mistura em um percentual acumulado que coincide com as quantidades das duas populações presentes. Na figura 11, por exemplo, um ponto de inflexão que aparece no percentil 15 indica 15% de uma população lognormal **A** (parte superior) e 85% da parte inferior de uma população lognormal **B**.

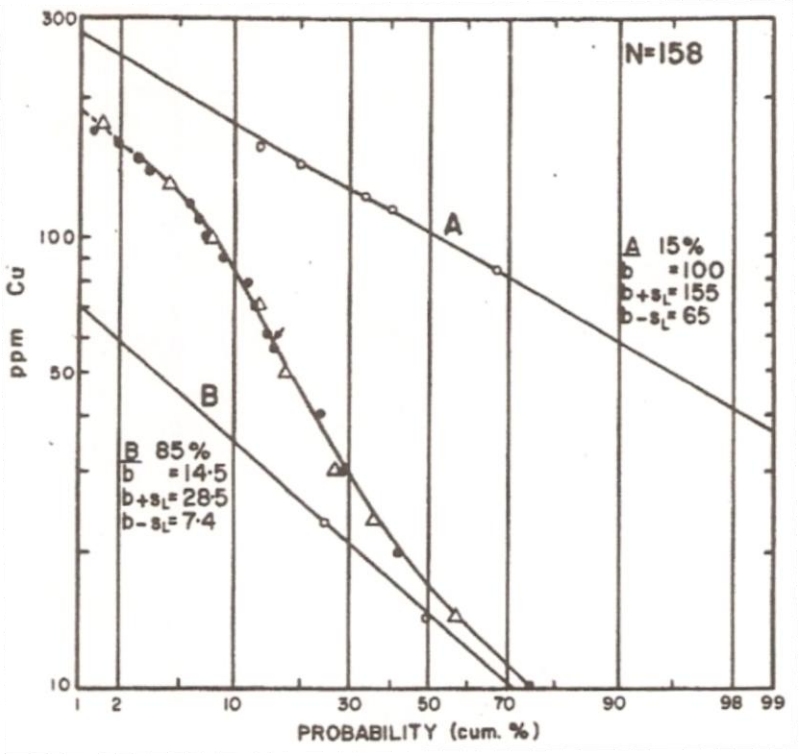


Figura 11 - Gráfico de Probabilidades de Sinclair

Pontos pretos são dados percentuais brutos acumulados, círculos abertos são pontos de construção que fornecem estimativas das populações **A** e **B** por partição da curva dos dados brutos, triângulos abertos são combinações calculadas das populações ideais **A** e **B** e fornecem uma checagem de quão bem o modelo de partição se ajusta à curva real. Uma pequena seta no percentil acumulado 15 mostra a posição estimada de um ponto de inflexão na curva dos dados brutos (de acordo com Saager e Sinclair, 1974).

PROPOSIÇÃO DE LEPELTIER

Lepeltier (1969) definiu um método de avaliação gráfica de somas cumulativas em escala gráfica bilogarítmica, em que apenas a concentração média do elemento (o valor do Clarke) é exigido. O método é baseado na pressuposição de que os valores dos elementos traços mostram distribuição lognormal.

ESTATÍSTICA NÃO PARAMÉTRICA

Estatística não Paramétrica é um conjunto de métodos que não utiliza os parâmetros da Distribuição Normal (média e desvio padrão). Os métodos não paramétricos são válidos independentemente do tipo e do tamanho das distribuições. Devido à falta de sentido da média para sua utilização, é utilizada a mediana; existem alternativas não paramétricas para todos os testes paramétricos padrão, e os cálculos para os testes não paramétricos são

baseados no *rank* dos dados. Os testes não paramétricos são referidos na bibliografia como robustos. Os testes não paramétricos (denominação mais utilizada) são chamados de testes livres de distribuição (que é uma denominação mais precisa), por não exigirem suposições sobre a natureza das distribuições envolvidas. Entre as vantagens dos métodos não paramétricos se pode citar sua ampla aplicabilidade, por não exigirem populações normalmente distribuídas e envolverem cálculos mais simples que os paramétricos correspondentes, e as desvantagens principais são a tendência a perder informação, pois os dados numéricos exatos são freqüentemente reduzidos a uma forma qualitativa, e não serem tão eficientes quantos os paramétricos; em um teste não paramétrico em geral se necessita de evidência mais forte (como uma amostra maior ou diferenças maiores) para se rejeitar uma hipótese nula. Quando são satisfeitas as exigências de distribuições populacionais os testes não paramétricos são geralmente menos eficientes que os seus correspondentes paramétricos, mas a redução da eficiência pode ser compensada por um aumento do tamanho da amostra; assim, se satisfeitas as condições exigidas é melhor que se utilize os testes paramétricos.

EDA e BOXPLOT

No final dos anos 1970, Tukey (1977) introduziu o paradigma de EDA - *Exploratory Data Analysis* - para analisar dados que não seguem um modelo normal. A EDA dá ênfase à exploração original, com o objetivo de simplificar a descrição dos dados e obter uma visão mais profunda da sua natureza. A EDA explora os dados em nível preliminar, poucas (ou talvez nenhuma) hipóteses são feitas sobre os dados, e costuma exigir cálculos e gráficos relativamente simples. Os diagramas têm a vantagem de não serem tão sensíveis a valores extremos como outras medidas baseadas na média e desvio padrão, e os *boxplots* não dão informação tão detalhada como os histogramas ou os gráficos ramo e folhas; ao utilizarmos diagramas em caixa para comparar dois ou mais conjuntos de dados é importante utilizarmos a mesma escala, de forma a possibilitar a comparação. EDA não é um método, mas uma filosofia ou uma abordagem robusta, consiste de um conjunto de estatísticas descritivas e principalmente ferramentas gráficas que visam a ganhar o máximo de informações sobre os dados, ajudar a descobrir a estrutura dos dados, definir variáveis significativas nos dados, determinar *outliers* e anomalias, que são um dos objetivos principais da interpretação de dados de prospecção geoquímica, sugerir e testar hipóteses de trabalho relativas a eventos geológicos, em especial os mineralizantes, e fornecer subsídios para a interpretação dos dados. A EDA usa estatística descritiva e gráficos robustos que são quantitativamente distintos dos da Estatística Clássica, os produtos gerados na EDA são baseados nos próprios dados e não em modelos de distribuição, como a normal, por exemplo.

BOXPLOT

Utilizamos, na descrição detalhada deste método, a terminologia original em inglês para que o leitor se familiarize com os termos da farta bibliografia em língua inglesa sobre o tema.

IQR - *Inter Quartile Range* é a diferença entre o terceiro e o primeiro quartis, e é uma medida de variabilidade que se concentra nos valores próximos ao centro da distribuição, representa bem o espalhamento dos dados na região central da distribuição.

A *Boxplot* (*Boxplot* também é conhecido como diagrama de *Box* e *Whisker*) tem como principais vantagens apontadas na bibliografia especializada a apresentação da distribuição dos dados e sua estatística, como locação central, assimetria, *outliers*, ser um método robusto de definição do *cut-off* para *outliers* que podem afetar parâmetros estatísticos na análise clássica, como média, desvio padrão e outros, e não assumir nenhum modelo particular estatístico para ajuste dos dados, evitando a necessidade de transformação dos mesmos. A caixa central inclui 50% dos dados centrais, as *whiskers* mostram a amplitude dos dados, isto é, a diferença entre os valores máximo e mínimo; a simetria é indicada pela caixa e pelos *whiskers* e pela localização da tendência central.

- A distância do Q1 à mediana é igual à distância da mediana até a Q3, a distância do mínimo até Q1 é igual à distância do valor máximo até Q3 para distribuições simétricas. Para valores assimétricos à direita a distância de Q3 até o valor máximo excede em muito a distância do valor mínimo até Q1, e a mediana é maior que a moda.
- Para conjuntos assimétricos à esquerda a distância do valor mínimo até Q1 excede em muito a distância de Q3 até o valor máximo, e a mediana é menor que a moda.

Os valores que se distanciam do restante do conjunto de dados se chamam dispersos ou *outliers*, que podem ser identificados como segue:

- a partir do valor mínimo se marca 3 x (Q3 - Q1), e todo valor entre o mínimo e este valor é um *outlier*
- a partir do valor máximo se marca 3 x (Q3 - Q1), todo valor que ficar entre o máximo e este valor é considerado um *outlier*
- alguns autores consideram que valores além da extremidade de 3 x (Q3 - Q1) são considerados dispersos extremos

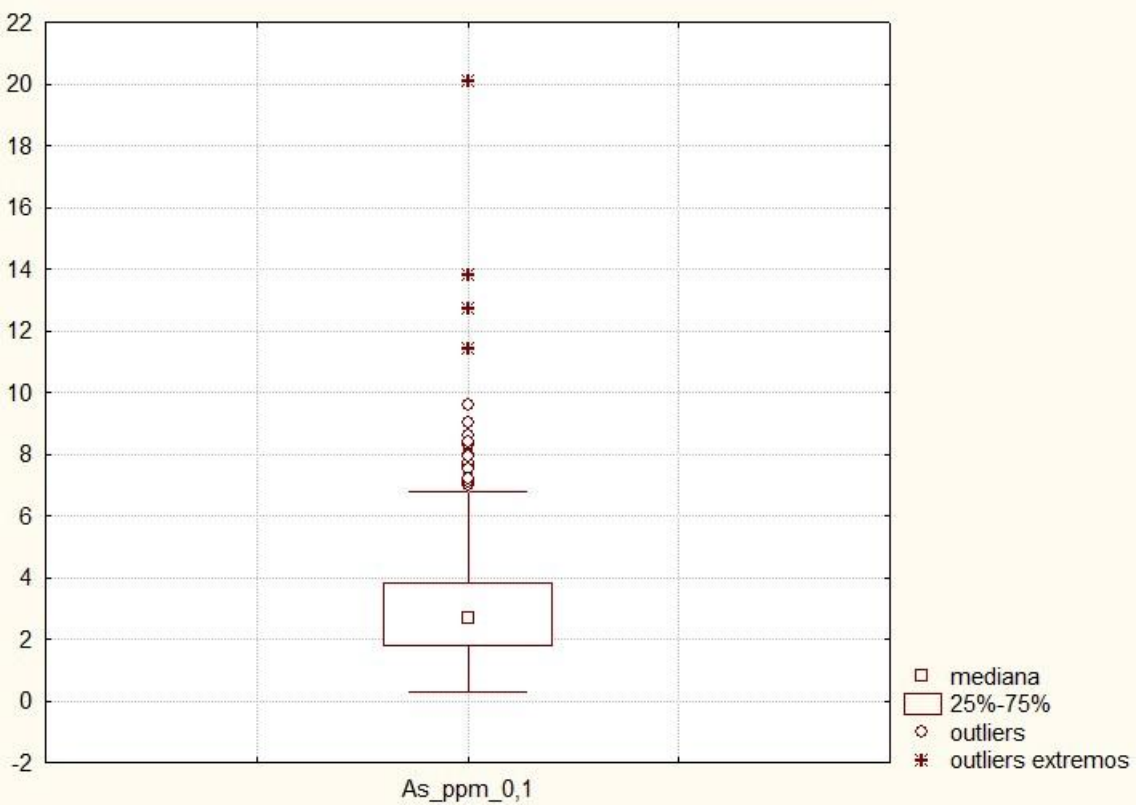


Figura 12 - Exemplo de Boxplot

DETALHAMENTO DO MÉTODO

Uma solução com muitas vantagens é utilizar o *box-plot* para a identificação de valores extremos. Os quartis superior e inferior, frequentemente referidos como *hinges*, definem a caixa central, que contém aproximadamente 50% dos dados. O *inner fence* é definido como uma caixa extendida por 1,5 vezes o comprimento da caixa em direção ao máximo e ao mínimo. Os valores que estão nos extremos das *inner fences* são os *whiskers*. Nas simulações lognormais as *fences* são calculadas usando os logaritmos dos valores, depois retrotransformados. Quaisquer valores fora do intervalo das *whiskers* são definidos como *outliers*. Valores além de três vezes (para mais ou para menos) a largura da *hinge* acima ou abaixo da superior ou da inferior são definidos como *far outliers*, isto é, valores não usuais para o conjunto de dados.

Pela contagem da metade da distância entre o mínimo e a mediana e entre o máximo e a mediana se define o ***lower hinge*** **(LH)** e o ***upper hinge*** **(UH)** respectivamente, estes três valores, **LH**, mediana e **UH** dividem os dados em quatro partes aproximadamente iguais denominadas quartis. A diferença absoluta entre **LH** e **UH** representa o *Inter Quartile Range* **(IQR)** ou a largura da *hinge*:

largura da ***hinge*** = **IQR** = $\left|$ ***lower hinge - upper hinge*** $\right|$

Define-se a LIF (*lower inner fence*) e a LOF (*lower outer fence*) respectivamente como valendo 1,5 x IQR e 3 x IQR além da LH em direção ao valor mínimo. Algebricamente elas valem (**X** representando seus valores numéricos)

$$XLIF = XLH - (1{,}5 \times IQR)$$

$$XLOF = XLH - (3 \times IQR)$$

Também se definem UIF (*upper inner fence*) e UOF (*upper outer fence*) como valendo respectivamente 1,5 x IQR e 3 x IQR além da *upper hinge* na direção do valor máximo, assim definidos algebricamente:

$$XUIF = XLH + (1{,}5 \times IQR)$$

$$XUOF = XLH + (3 \times IQR)$$

Uma LW (*lower whisker*) e uma UW (*upper whisker*) são plotadas a partir de cada uma das *hinges* em direção aos dados mais extremos dentro das *inner fences*, algebricamente os valores **X** da **LW** e da **UW** podem ser determinados como segue:

$$XLW = \min(X[X > XLIF])$$

e

$$XUW = \max(X[X < XUIF])$$

onde os valores entre colchetes são aqueles que estão dentro das *inner fences* e as *hinges*. Dados fora das *inner fences* são considerados *outliers*, dados entre a *inner* e a *outer fence* são considerados *mild* (suave, brando) *outliers*, enquanto os dados fora das *outer fences* são considerados *far* ou *extreme outliers*, isto é, valores muito não usuais. *Mild* e *extreme outliers* devem ser marcados por símbolos diferentes.

A *boxplot* (ou *Box-and-Whisker*) define cinco estatísticas sumárias, o mínimo, a **LH**, a mediana, a **UH** e o máximo, e descreve as características mais importantes de um conjunto de dados, ou seja, sua tendência central, seu espalhamento, sua assimetria, seus comprimentos de caudas e seus *outliers*, e é resistente com relação à presença de *outliers* nos dados. Com base em um *boxplot* uma exploração unielementar de dados geoquímicos nos permite dividir os valores em cinco classes robustas, que são

- **mínimo - LW**
- **LW - LH**
- **LH - UH**
- **UH - UW**
- **UW - máximo**

A **UIF** é usualmente considerada o *threshold* separando valores de *background* e anomalias, embora a **UOF** também possa ser usada como *threshold*.

Assim, valores na classe (UH - UW), pelo menos 25% dos dados de um conjunto de valores, pode ser considerada como alto *background*, valores entre (LH - UH), classe com cerca de 50% dos dados, representa o *background*, e valores entre (LW - LH), classe com até 25% dos dados, formam o baixo *background*, e valores na classe (mínimo - LW) formam o *background* extremamente baixo. À parte do *threshold* definido pela *boxplot* (como **UIF** ou **UW**), um *threshold* pode ser definido a partir da EDA como (mediana + 2 MAD). O MAD é similar ao desvio padrão da estatística clássica, de modo que esta abordagem se assemelha à abordagem de (média + 2 desvios padrão).

A simbologia utilizada pelo *boxplot* pode ser utilizada para compor a legenda de mapas, pois tem intervalos robustos. A classificação de dados geoquímicos com base na EDA e nas classes da *boxplot* tem a forte habilidade de representar e dar significado físico às distribuições unielementares sem a necessidade de adoção da normalidade das distribuições ou informações prévias sobre certos fatores que influenciam a variabilidade de um conjunto de dados geoquímicos. É importante citar que uma vantagem desta representação é a simplificação em termos de quantidades de classes, pois não há muita utilidade prática em se dividir uma representação gráfica em uma grande quantidade de classes com símbolos ou cores diferentes quando o que se busca, na realidade, são poucas categorias que tenham, cada uma, um significado geoquímico específico (*background*, *outlier*).

Reinmann et al. (2005) colocam que o valor de *threshold* (UW) da *boxplot* é adequado em casos onde se tem menos de 10% de outliers, enquanto que a abordagem (mediana + 2 MAD) é adequada nos casos em que se tem pelo menos 15% de outliers.

A tabela 6 mostra comparações entre as diferentes determinações de *threshold* utilizando algumas técnicas.

	MÉDIA	MEDIANA	I QUARTIL	III QUARTIL	IQR	DESV. PADRÃO	MÉDIA +(2 DESVPAD)	MG*DG*DG	BOXPLOT	MEDIANA + (2MAD)
Ag_ppb_2	17,87	13,00	10,00	17,00	7,00	46,33	110,52	39,39	27,50	21
As_ppm_0,1	3,05	2,70	1,80	3,80	2,00	1,89	6,83	8,41	6,80	4,7
Au_ppb_0,2	5,24	0,60	0,30	0,90	0,60	85,67	176,58	4,75	1,80	1,2
Ba_ppm_0,5	106,70	101,50	75,40	126,20	50,80	48,55	203,80	237,48	202,40	152,1
Bi_ppm_0,02	0,28	0,27	0,21	0,33	0,12	0,20	0,67	0,79	0,51	0,39
Co_ppm_0,1	10,41	9,50	6,80	13,00	6,20	5,72	21,85	28,56	22,30	15,5
Cu_ppm_0,01	12,33	11,51	8,07	15,13	7,06	6,87	26,06	33,02	25,72	18,65
Hg_ppb_5	39,33	36,00	28,00	46,00	18,00	17,06	73,46	80,31	73,00	54
Mo_ppm_0,01	0,49	0,22	0,16	0,32	0,16	3,24	6,96	0,98	0,56	0,36
Ni_ppm_0,1	10,61	9,30	6,10	13,30	7,20	7,80	26,21	33	24,10	16,3
Pb_ppm_0,01	22,42	20,73	16,40	25,20	8,80	12,40	47,22	44,79	38,40	29,49
Sn_ppm_0,1	0,98	0,80	0,60	1,00	0,40	2,64	6,26	1,98	1,60	1,2
Zn_ppm_0,1	48,97	46,80	33,90	62,00	28,10	22,88	94,73	130,98	104,15	74,2
P_%_0,001	0,03	0,03	0,02	0,04	0,02	0,03	0,09	0,08	0,08	0,04

	< (Q1 - 1,5 IQR)	(Q1 - 1,5 IQR) a MEDIANA	MEDIANA a Q3	Q3 a (Q3 + 1,5 IQR)	> (Q3 + 1,5 IQR)
	FAIXA 1	FAIXA 2	FAIXA 3	FAIXA 4	FAIXA 5
Ag_ppb_2		0 a 13	13 a 17	17 a 27,5	> 27,5
As_ppm_0,1		0 a 2,7	2,7 a 3,8	3,8 a 6,80	> 6,80
Au_ppb_0,2		0 a 0,60	0,6 a 0,9	0,9 a 1,80	> 1,80
Ba_ppm_0,5		0 a 101,50	101,5 a 126,20	126,20 a 202,4	> 202,4
Bi_ppm_0,02	< 0,03	0,03 a 0,27	0,27 a 0,33	0,33 a0,51	> 0,51
Co_ppm_0,1		0 a 9,50	9,5 a 13,00	13,00 a 22,30	> 22,30
Cu_ppm_0,01		0 a 11,51	11,51 a 15,13	15,13 a 25,72	> 25,72
Hg_ppb_5	< 1	1 a 36	36 a 46	46 a 73	> 73,00
Mo_ppm_0,01		0 a 0,22	0,22 a 0,32	0,32 a 0,56	> 0,56
Ni_ppm_0,1		0 a 9,30	9,30 a 13,30	13,30 a 24,10	> 24,10
Pb_ppm_0,01	< 3,2	3,2 a 20,73	20,73 a 25,20	25,20 a 38,40	>38,40
Sn_ppm_0,1		0 a 0,80	0,80 a 1	1 a 1,60	> 1,60
Zn_ppm_0,1		0 a 46,80	46,80 a 62	62 a 104,15	> 104,15
P_%_0,001		0 a 0,03	0,03 a 0,04	0,04 a 0,07	> 0,07

Tabela 6 - Determinações de *Threshold*

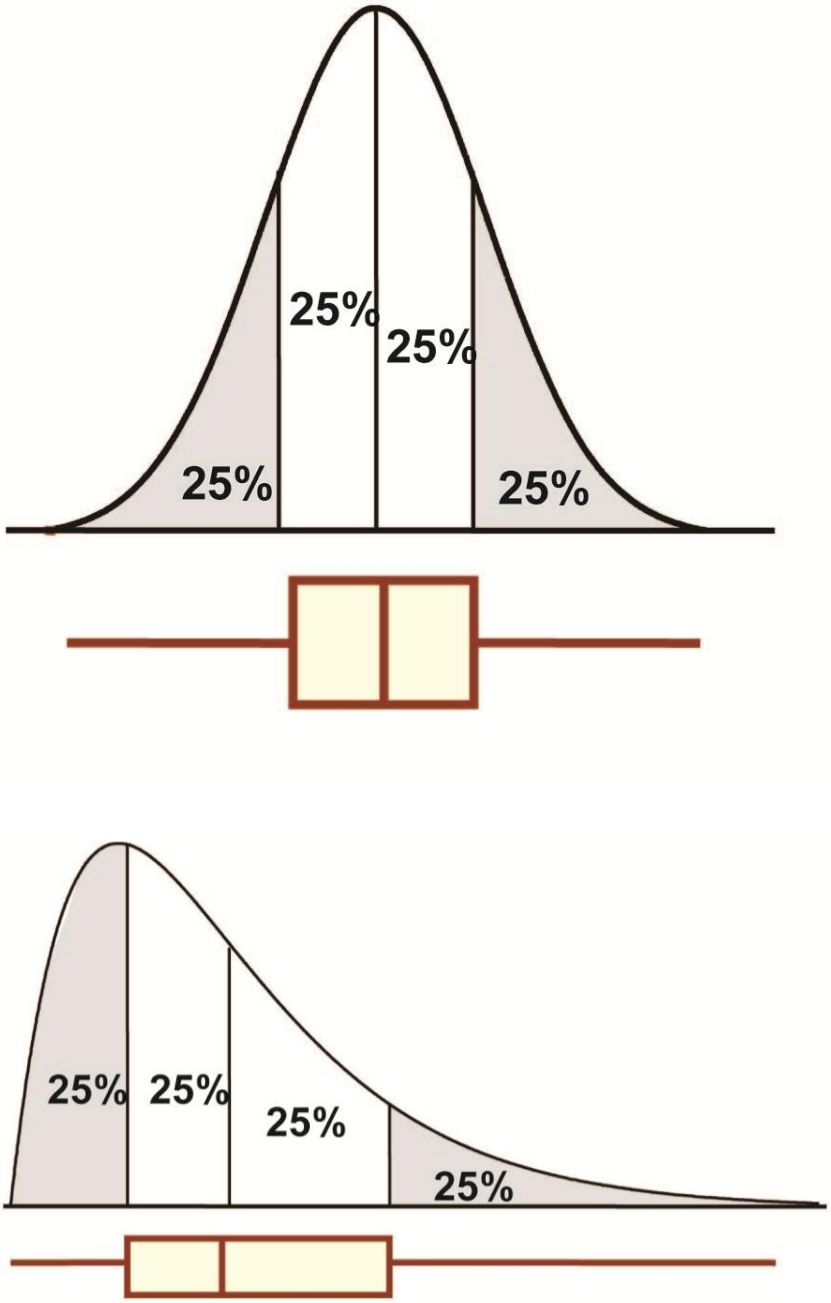


Figura 13 - Boxplot

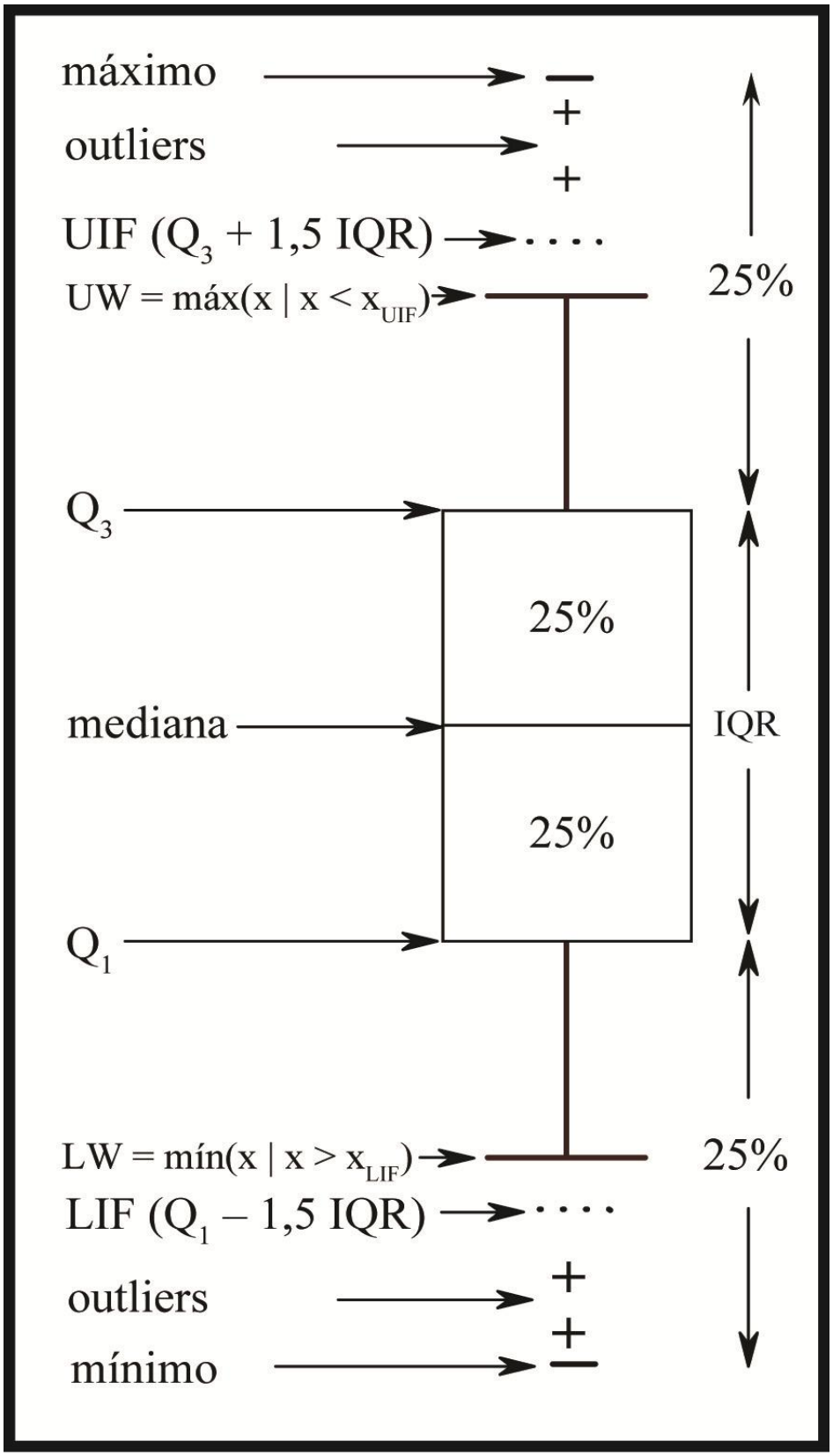


Figura 14 - Boxplot - Esquema

ESTATÍSTICA MULTIVARIADA

Em termos de estatística multivariada, as anomalias combinadas podem ser mais robustas ou indicativas de um tipo particular de fonte que as anomalias de elementos individuais; por exemplo, anomalias associadas com EGP podem ser associadas para discriminar entre anomalias de níquel geradas a partir de depósitos sulfetados de níquel hospedados em komatiitos e anomalias relacionadas ao intemperismo de rochas ultramáficas portadoras de níquel. Na prospecção geoquímica se trabalha com muitas variáveis simultaneamente, pois que é o conjunto de variáveis que modela uma paisagem geoquímica, e não apenas uma isoladamente. Como as diversas variáveis interagem para formar o quadro final observado, algumas destas interações e associações aparecem, às vezes, de forma clara nos estudos multivariados. É importante ressaltar que os resultados obtidos nas aplicações multivariadas devem, como em todos os casos de interpretação de dados, ser confrontados com as informações geológicas disponíveis na área de estudo. Deve-se buscar, para cada resultado de uma aplicação multivariada, uma associação com um processo geológico, metalogenético ou geoquímico da área. Assim, ao se representar um resultado desses em um mapa, estaremos representando o comportamento de um processo que atuou na região, o que pode ser de grande valia no entendimento da evolução da mesma, especialmente no caso de representar mineralização, alteração hidrotermal ou outro. No presente trabalho abordamos a Análise de Agrupamentos (*Cluster Analysis*) e as técnicas de Análise Fatorial e Análise de Componentes Principais.

ANÁLISE DE CONGLOMERADOS (*CLUSTER ANALYSIS*)

Esta técnica tem por objetivo classificar objetos (no caso da geoquímica, amostras de campo ou variáveis analisadas) por similaridades e/ou dissimilaridades, os grupos gerados devem ter alta homogeneidade interna e alta heterogeneidade externa. Não há nenhuma definição prévia com relação ao número de grupos ou sua estrutura, cabendo à teoria a definição prévia deste número. É uma técnica comumente utilizada para formular hipóteses sobre a natureza dos dados ou para examinar hipóteses já existentes sobre os mesmos. A Análise de Conglomerados não é uma técnica de inferência estatística (tem propriedades matemáticas fortes mas não tem fundamentos estatísticos, ou seja, não tem base estatística sobre a qual se possa formular inferências sobre uma população com base em uma amostra, suas soluções não são únicas e sempre são criados grupos, quer eles existam na estrutura dos dados ou não). A presença de *outliers* nos dados afeta os resultados da Análise de Agrupamentos. É importante ter em mente que os grupos gerados podem ser naturais, gerados por associações entre as variáveis disponíveis, ou mesmo artificiais, pois que a técnica busca formar grupos, e eles podem não ter significado geológico em alguns casos.

Os métodos mais comuns são os que utilizam medidas de correlação (o coeficiente de correlação é uma medida de similaridade), medidas de distância (distância euclidiana é uma medida de dissimilaridade, altos valores representam menores similaridades, uma medida desta distância que incorpora diretamente o procedimento de padronização dos dados se chama Distância de Mahalanobis, $\mathbf{D^2}$) e de associação, as duas primeiras exigem dados quantitativos, a última dados qualitativos. É necessário ter cuidado com as escalas das variáveis, escalas diferentes influenciam no resultado final, deve-se padronizar as variáveis sempre que possível (variáveis com altas dispersões afetam os resultados obtidos).

Os algoritmos hierárquicos apresentam resultados similares a uma árvore, e a figura resultante se chama dendrograma. Dentre os mais utilizados estão o *Single Linkage* ou *Nearest Neighbour*, que utiliza a distância mínima, dois objetos separados pela menor distância ficam no mesmo grupo, depois é agregado a este grupo o objeto que estiver à menor distância deste grupo, e assim por diante, sendo a distância entre dois grupos a menor entre qualquer ponto de um grupo e qualquer ponto de outro grupo. Tende a gerar grupos mais equilibrados e menos díspares.

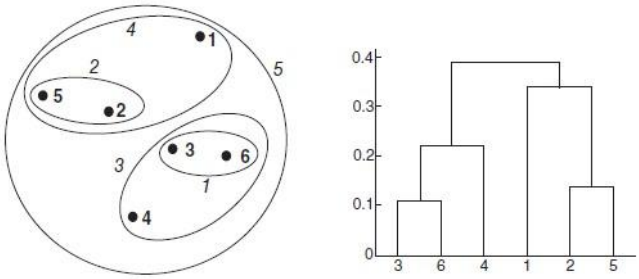
O *Complete Linkage* ou *Furthest Neighbour* se diferencia do *Single Linkage* por agrupar os objetos pela distância máxima.

O *Average Linkage* ou *Between-Groups Linkage* e *Within-Groups Linkage* adotam a distância entre todos os indivíduos de um grupo e todos os indivíduos do outro grupo, favorece a formação de grupos com menores variações internas.

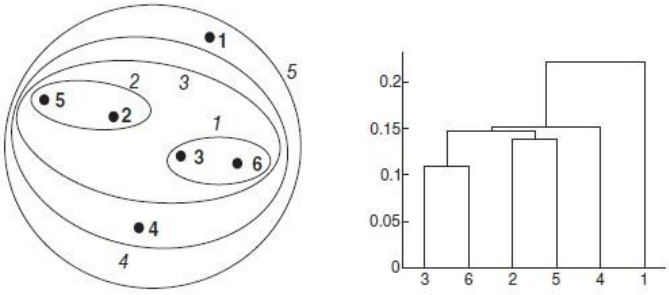
O Método de *WARD* utiliza a perda de informação que ocorre quando da formação dos grupos, utiliza a soma total dos quadrados dos desvios de cada objeto em relação à média do conglomerado em que o objeto é classificado, este valor vai sendo reduzido gradativamente à medida que as interações avançam, os grupos geralmente são menores em tamanho.

É uma prática recomendável utilizar mais de um algoritmo de formação de grupos e confrontar os resultados, concluindo-se que grupos que aparecem de forma similar em vários algoritmos sejam mais consistentes que grupos que aparecem em poucos ou em apenas um dos métodos de agrupamento escolhidos. Em outras palavras, o agrupamento é de tal forma destacado que independe do algoritmo para se manifestar. A aplicação da técnica de Análise de Agrupamentos também é muito útil como ferramenta exploratória dos dados, ou seja, ela pode indicar associações/agrupamentos de que não suspeitávamos previamente, como também pode, muitas vezes, simplesmente refletir algumas associações litológicas óbvias, como Ni/Co/Cr em rochas ultramáficas, Pb/Zn, Au/Ag/As/Bi, EGP, ETR, Ba/Sr e muitas outras em diversas litologias.

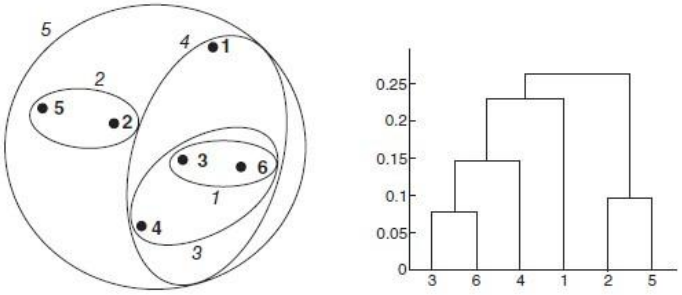
A figura 15 mostra um exemplo em que dados são agrupados por três algoritmos diferentes, gerando grupos diferentes.



complete linkage



single linkage



método de Ward

Figura 15 - Dendrograma Três Algoritmos

Nas figuras 16, 17 e 18 são apresentados resultados de uma aplicação da Análise de Agrupamentos a dados litológicos em área de bacias Eoproterozóicas-Neoproterozóicas do sul do Brasil. Foram feitas aplicações desta técnica por três algoritmos diferentes (*Single Linkage, Complete Linkage*, Método de Ward), notando-se o aparecimento dos mesmos grupos de amostras nas três aplicações (amostras C 42, 40, 39, 27, 18, 20, 17, 15, 13, 11 e 10 num grupo e as demais em outros grupo), aparecendo diferenças apenas ao se detalhar mais a interpretação e a leitura dos dendrogramas gerados. Em outras palavras, os grandes agrupamentos presentes se mostraram consistentes quando os resultados analíticos foram tratados por diferentes métodos.

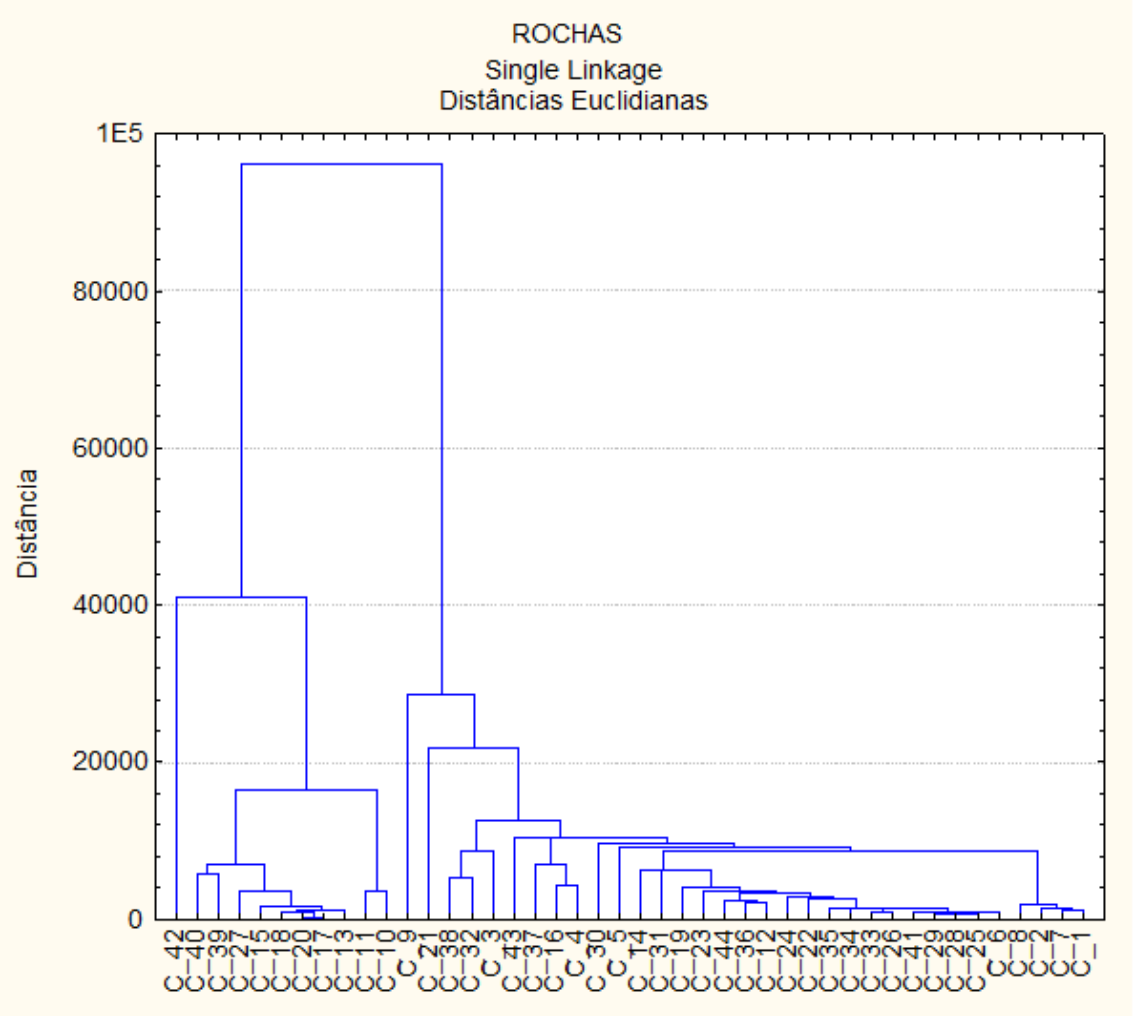


Figura 16 - Dendrograma *Single Linkage*

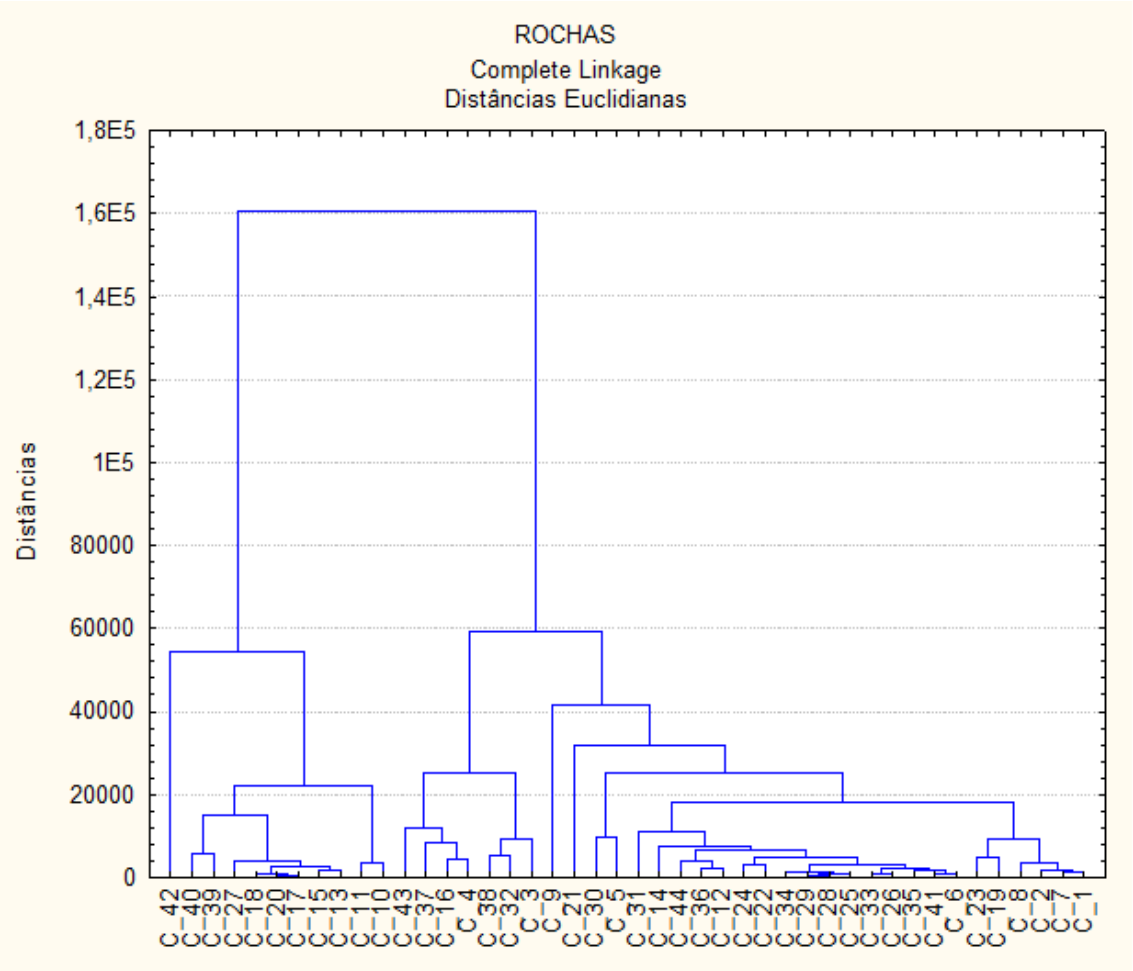


Figura 17 - Dendrograma *Complete Linkage*

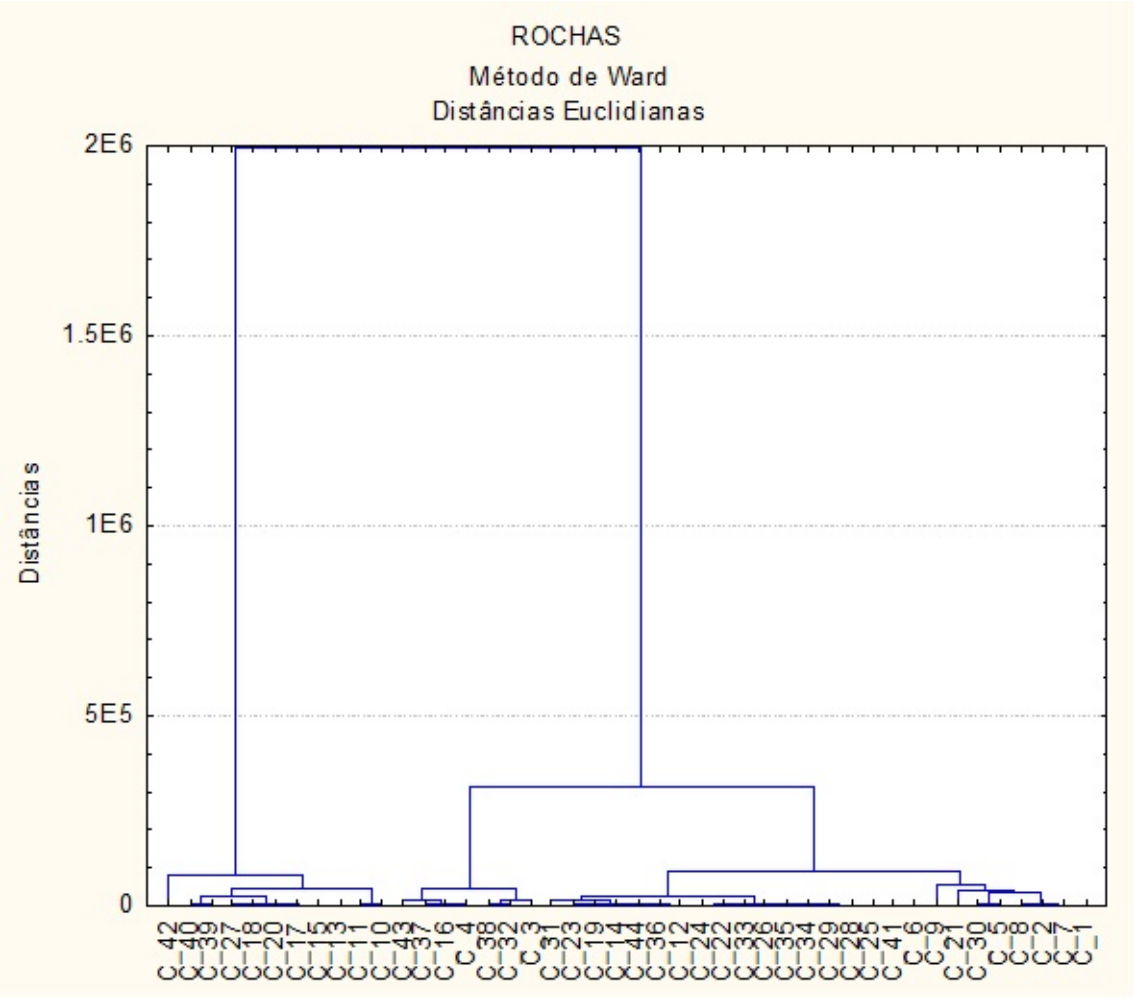


Figura 18 - Dendrograma Método de Ward

ANÁLISE FATORIAL (AF) E ANÁLISE DE COMPONENTES PRINCIPAIS (ACP)

Na **Análise Fatorial** se calcula a matriz de correlação para todos os pares possíveis de variáveis, a matriz é diagonalizada e suas componentes principais, os *eigenvectors* (referidos como autovetores na língua portuguesa), são obtidos; o primeiro *eigenvector* será relacionado ao maior *eigenvalue* (referidos como autovalores na língua portuguesa) e explicará a maior quantidade possível da variância dos dados, o segundo *eigenvector* será ortogonal e não correlacionado ao primeiro, e assim por diante. *Eigenvectors* são vetores linearmente independentes que são combinações lineares das variáveis originais, e podem ser vistos como “novas” variáveis com a desejável propriedade de serem não correlacionadas e responderem pela variância dos dados em ordem decrescente de importância.

Na Análise Fatorial eles recebem o nome de fatores. Comunalidades das variáveis são a proporção da sua variância explicada pelos fatores comuns extraídos, cada variável pode ser expressa na forma **x = a . f**, onde a indica **a** importância relativa da componente e **f** indica o fator comum; muitos problemas são relacionados apenas ao fator comum, a soma dos quadrados dos coeficientes dos fatores comuns responde pela variância daquela variável, é a comunalidade. O modelo de Análise Fatorial assume o conhecimento do número de fatores. Na prática este número é muitas vezes desconhecido e valores diferentes podem ser tentados seqüencialmente, começando com o valor unitário. Não é fácil, entretanto, selecionar este valor corretamente. É desagradável verificar que a interpretação dos fatores possa mudar completamente com a mudança deste número.

Se os perfis são plotados pelos escores de suas duas primeiras componentes o plano resultante é o melhor sumário bidimensional do espaço **r-**dimensional. Em certos casos é aplicada a rotação *Varimax,* um critério matemático utilizado para maximizar a variância.

Variável	CP1	CP2	CP3	Comunalidade
A	**0,94**	**0,13**	**-0,02**	**0,91**
B	**0,55**	**0**	**0,31**	**0,40**
C	**-0,04**	**0,02**	**0,93**	**0,87**
D	**-0,04**	**0,94**	**0,02**	**0,89**
E	**-0,03**	**0,95**	**0,03**	**0,91**
F	**0,23**	**0,81**	**0,09**	**0,72**
G	**0,93**	**0,03**	**-0,04**	**0,87**
H	**0,94**	**0,08**	**0,03**	**0,89**
I	**0,25**	**0,61**	**0,03**	**0,80**

A coluna comunalidade indica qual a proporção da variabilidade de cada variável particular explicada pelos três fatores mostrados, é obtida pelos quadrados dos valores absolutos em cada fila, por exemplo a comunalidade de 0,91, da variável **A**, é obtida pela soma dos quadrados de 0,94, 0,13 e –0,02. A baixa comunalidade da variável **B** sugere que ela provavelmente estaria melhor contida em um espaço de dimensionalidade maior. A comunalidade reflete o grau ao qual cada amostra/vetor foi explicada pelo conjunto de eixos, e comunalidade igual a um indica uma perfeita explicação.

São sinônimos *eigenvalues*, *latent roots*, *characteristic roots*, *proper values*, bem como *eigenvectors* com os mesmos termos trocando-se *roots* por *vectors*; associado a cada *eigenvalue* há um *eigenvector* que é único, exceto para o seu sinal, *eigenvectors* associados com diferentes *eigenvalues* são ortogonais.

A **Análise de Componentes Principais** transforma **p** variáveis originais correlacionadas em **q** variáveis não correlacionadas, sendo **q < p**, ou seja, em **q** componentes principais que sejam funções lineares das variáveis originais, e esta transformação se dá de tal modo que a projeção das observações originais no novo sistema seja nula, ou seja, projeção ortogonal; assim, o eixo original (variável **p**) é transformado em novo eixo, a componente principal, que é estatisticamente independente, pois que cada componente principal, por sua vez, é ortogonal à seguinte.

A **Análise de Componentes Principais** é uma técnica matemática que não requer que o usuário especifique um modelo estatístico subjacente para explicar o erro. Em particular, nenhuma condição é imposta sobre a distribuição de probabilidade das variáveis originais. Cada nova variável (componente

principal, CP) gerada é uma combinação linear das **p** variáveis originais, podendo ser escrita como:

$$\mathbf{CP_i = A_{i1}X_1 + A_{i2}X_2 + ... + A_{ij}X_j + ... + A_{ip}X_p}$$

onde **i = 1, ... , p**, em que $\mathbf{X_1, X_2, ... , X_p}$ são as variáveis originais e em que os $\mathbf{A_{ij}}$ são coeficientes. As componentes principais são tais que:

i) Variância CP_1 > Variância CP_2 > ... > Variância CP_q

ii) Os valores de quaisquer duas CP são não correlacionados, o que não ocorre com as variáveis originais

iii) É usual reter apenas as primeiras **q** componentes **(q << p).**

Analisemos, agora, a matriz de autovetores abaixo: $\mathbf{X_1, X_2,...,X_p}$ são as variáveis originais disponíveis, **e** $\mathbf{CP_1, CP_2,...,CP_p}$ são as componentes principais obtidas (esta matriz é denominada matriz dos coeficientes das componentes principais).

	$\mathbf{CP_1}$	$\mathbf{CP_2}$.....$\mathbf{CP_p}$
$\mathbf{X_1}$	λ_{11}	λ_{12}.....λ_{1p}
$\mathbf{X_2}$	λ_{21}	λ_{22}.....λ_{2p}
$\mathbf{X_p}$	λ_{p1}	λ_{p2}.....λ_{pp}

Na mesma matriz acima, λ_{ij} representa o coeficiente de correlação linear entre a variável $\mathbf{X_i}$ e a componente principal $\mathbf{CP_j}$. A soma dos quadrados dos valores de toda uma coluna desta matriz, λ_j para a raiz característica associada à $\mathbf{CP_j}$ (esta raiz característica representa a variância da componente principal), e o valor de λ_j dividido por **p** representa a proporção da variância das variáveis originais explicada pela j-ésima componente principal.

Se para um determinado caso a comunalidade para uma determinada variável for 0,80, por exemplo, significa que 80% da variabilidade daquela variável é explicada pelas componentes principais; a diferença de 20% é denominada especificidade, e representa a parte da variabilidade da variável em questão que não é explicada por nenhuma das componentes principais. Há determinados pontos que devem ser analisados com cuidado ao se utilizar a Análise de Componentes Principais sobre um conjunto de dados. Um deles é que se deve evitar a utilização de variáveis que sejam combinações de outras

variáveis do mesmo conjunto, pois que isto pondera variáveis em relação a outras, o que não é desejável. Outra situação contra indicada por grande parte dos autores é a aplicação de técnicas multivariadas a conjuntos de dados em que o número de variáveis seja maior que o número de observações disponíveis. A presença de valores aberrantes (*outliers*) em meio aos dados acarreta grandes danos à interpretação por encobrir relações significativas que possam estar presentes nos dados ou, mesmo, por ressaltar relações não significativas - adulteração, em suma, da realidade subjacente. Também não se deve aplicar a Análise de Componentes Principais para amostras em que o número de observações disponíveis seja pequeno.

Em contraste com a Análise Fatorial, as componentes derivadas da Análise de Componentes Principais são únicas (exceto quando houver autovalores iguais) e que permanecem as mesmas quando variamos o número de componentes que são pensados necessários incluir. Na literatura de aplicação de técnicas multivariadas muitos autores têm confundido as técnicas de Análise de Componentes Principais e Análise Fatorial, não considerando que a Análise de Componentes Principais, diferentemente da Análise Fatorial, não assume nenhum modelo matemático bem definido. Como resultado, pode-se encontrar exemplos na literatura geológica que alegadamente tratam de Análise Fatorial quando na realidade apenas Análise de Componentes Principais foi usada e vice-versa. **Em Análise Fatorial a adição de outro fator pode modificar os anteriores. Em Análise de Componentes Principais a adição de outra componente principal deixará as componentes principais anteriores inalteradas.**

As diferenças fundamentais entre ACP e AF dependem dos modos nos quais os fatores são definidos e das suposições que dizem respeito à natureza dos resíduos. Em ACP as componentes são determinadas de modo a responder pela maior parte da variância das variáveis estudadas, e na AF os fatores são definidos para responder maximamemte pelas interrelações das variáveis. Assim, ACP é dita ser orientada para a variância e a AF é dita ser orientada para a correlação. O resíduo é assumido ser pequeno na ACP, o que não ocorre na AF.

A abordagem *Varimax* de Kaiser é para encontrar um novo conjunto de posições para os fatores principais de modo a que a variância das cargas fatoriais ou coeficientes das componentes em cada fator seja um máximo; as cargas fatoriais tenderiam à unidade ou a zero. Os escores fatoriais para os fatores *Varimax* também podem ser computados, e os valores podem ser interpretados do mesmo modo que os escores dos fatores principais. O método de rotação ortogonal mais popular é o *Varimax*.

Rotação a uma solução não ortogonal é permissível, é a chamada rotação oblíqua. Fatores correlacionados são aceitáveis se eles explicam a estrutura.

ACP é uma forma de identificar padrões nos dados e expressá-los de modo a ressaltar suas similaridades e diferenças, com a vantagem de que uma vez identificados estes padrões podemos reduzir o número de dimensões sem muita perda de informação.

- Escolhe-se componentes e se forma uma feição vetorial, e o *eigenvector* com o maior *eigenvalue* correspondente é a componente principal do conjunto de dados. Pode-se ignorar as componentes de menor importância, com certa perda de informação, mas se os *eigenvalues* forem pequenos a perda não é muito significativa.
- O passo seguinte é o em que escolhemos as componentes (*eigenvectors*) que serão mantidas em nossos dados e que formarão o vetor da feição. O que isto nos dará? Nos dará os dados originais somente em termos dos vetores que nós escolhemos. Podemos expressar os dados em termos de quaisquer dois eixos escolhidos, sendo eles perpendiculares a expressão é a mais eficiente, é por esta razão que é importante que os *eigenvectors* sejam sempre perpendiculares entre si. O que fizemos foi, basicamente, transformar os dados de modo que fossem expressos em termos de padrões entre eles, onde os padrões são as linhas que mais proximamente descrevem as relações entre os dados, isto é útil porque nos permite classificar os dados como uma combinação das contribuições de cada uma destas linhas.
- Deve-se ter em mente que, ao retomar os dados originais de volta, somente se tomarmos todos os *eigenvectors* em nossas transformações nós obteremos exatamente os dados originais de volta. Se tivermos reduzido o número de *eigenvectors* nas transformações finais então os dados recuperados terão sofrido perda de alguma informação.

Existem várias técnicas para a tomada de decisão sobre a retenção de componentes principais, sendo a mais aceita e recomendada a que propõe a retenção de componentes principais cujos percentuais de variabilidade totais sejam considerados significativos, ou seja, que expliquem uma proporção importante das variações totais presentes no conjunto original de dados. É evidente que, neste ponto, é decisiva a participação do técnico que aplica a Análise de Componentes Principais; a não utilização de componentes a partir de um certo nível (a partir da quarta componente principal, ou da quinta, por exemplo) não causa, entretanto, nenhum problema sobre as utilizadas, nem afeta sua interpretação sob hipótese alguma. Pode, tão somente, deixar inexplicados processos representados pelas componentes não utilizadas, o que não tem nenhuma influência, como dito acima, sobre as componentes (e seus processos associados) interpretadas. Um outro procedimento muito utilizado e recomendado por muitos autores e que está presente como opção na grande maioria dos *softwares* disponíveis é o de reter apenas as componentes principais cujos autovalores tenham valor superior à unidade. Os programas de

computador têm, normalmente, a opção de executarem o gráfico em que no eixo das abcissas se coloca as componentes principais a partir da primeira e no eixo das ordenadas se colocam os valores dos autovalores correspondentes. Este é um procedimento muito bem aceito pela quase totalidade dos autores que tratam do assunto. Esta regra é o teste SCREE, devido a Cattell (1966); o ponto de quebra é o ponto de parada de retenção. Um problema deste método é que é subjetivo, e outro é que pode não haver quebra no gráfico, e pode, também, haver mais de uma quebra, quando é usual se adotar a primeira quebra. A figura 19 é um exemplo da gráfico citado.

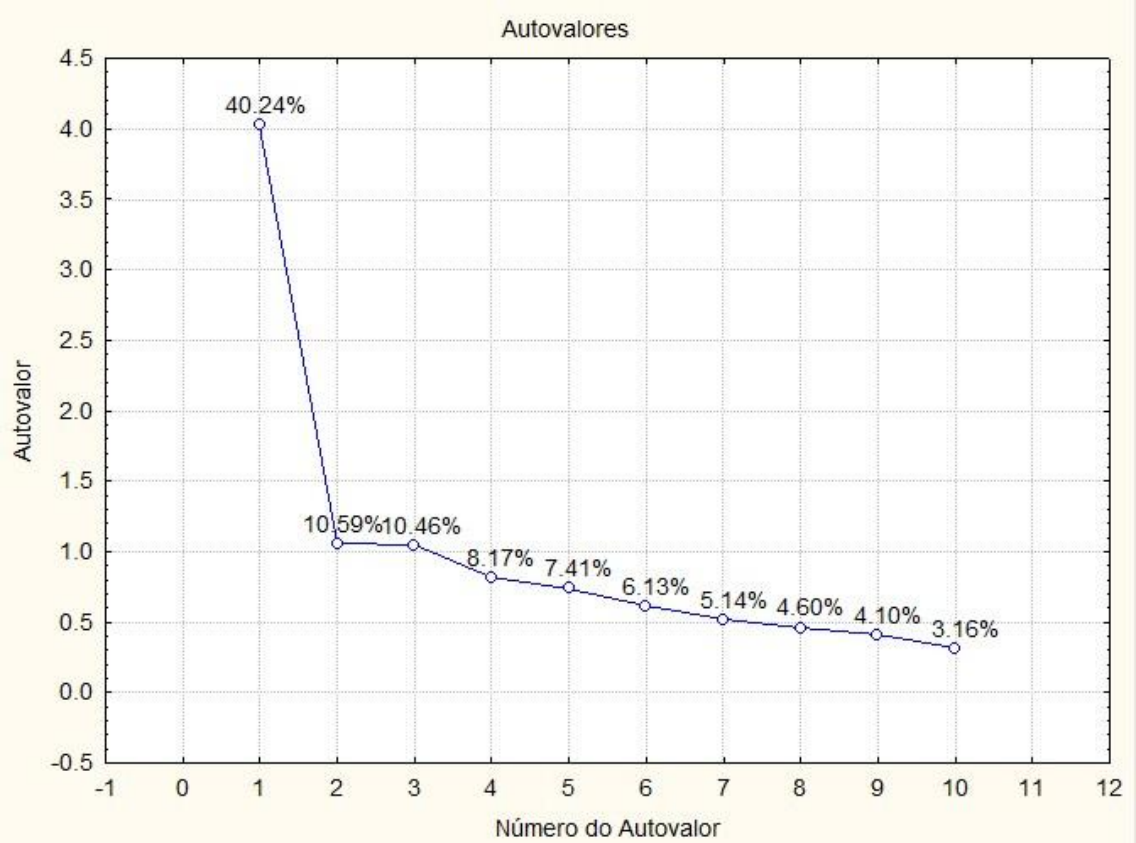


Figura 19 - Regra de Cattell (SCREE)

Entende-se que o procedimento mais indicado seria o de verificar todas as possibilidades acima, mas destacando o percentual explicado pelas componentes principais retidas e os valores numéricos dos autovalores superiores a um. A conjugação destes dois procedimentos é a prática mais utilizada pelos analistas.

Ao se utilizar as primeiras componentes principais (as que explicam maiores percentuais individuais de variabilidade) se pode representá-las em um sistema de eixos bi-dimensional ortogonal, marcando sobre o plano os escores das componentes principais um e dois para as diversas variáveis: é o melhor sumário bi-dimensional do espaço vetorial disponível de que se pode dispor. Também, o escore fatorial de cada componente para cada elemento da amostra pode ser calculado e os *softwares* disponíveis calculam estes valores.

Os eixos fatoriais são os valores de cada fator ou componente (variável hipotética) para cada elemento da amostra.

Em se tratando de variáveis regionalizadas, como é o caso em Geologia, em que a representação em mapa é de fundamental importância, a Análise de Componentes Principais representa ganhos enormes em a) termos de associação de variáveis e sua interpretação em relação aos processos geológicos atuantes, b) na simplificação da interpretação, pois que reduz a dimensionalidade do problema e enseja a diminuição de custos em etapas posteriores da investigação geológica e, principalmente, c) permite a visualização, em mapas, da distribuição de valores correspondentes a determinados eventos geológicos (as componentes principais), o que raramente é possível em mapas de uma única variável.

Com a Análise de Componentes Principais se identifica novas variáveis significativas, que, em Geologia, podem, por exemplo, representar processos geológicos atuantes, associações litológicas presentes ou eventos causadores de mineralização. Reduz-se, também, a dimensionalidade do problema, e, em adição, permite identificar variáveis originais que contribuem muito pouco para a elucidação do comportamento de uma área quando se estuda os processos que nela atuaram, identificação que permite sugerir a eliminação destas variáveis em etapas futuras de estudo.

Se algumas das variáveis originais são altamente correlacionadas elas estão, efetivamente, "dizendo a mesma coisa" e podem ser condições quase lineares sobre as variáveis.

Um objetivo muito comum em programas de investigação geológica é a classificação das observações que compõem uma determinada amostra, sejam estas de rochas, de sedimentos de corrente, de solos, de fósseis ou de água, por exemplo; esta classificação se dá com base nas propriedades composicionais medidas. É muito útil nos casos em que se dispõe de grandes quantidades de observações e, também, naqueles em que há pouco conhecimento prévio do significado da gênese dos constituintes. A Análise Fatorial Modo Q contempla este objetivo, permitindo a análise das inter-relações entre os espécimens disponíveis. Em termos matemáticos as técnicas Modo Q operam sobre matrizes de distâncias entre indivíduos. Quando a natureza de um problema geológico é tal que as relações entre entidades são o foco, então o Modo Q se torna uma ferramenta útil. A Análise Fatorial Modo Q pode ser vista como uma técnica de agrupamento. O delineamento de litofácies ou biofácies são talvez as situações mais evidentes deste tipo, onde o objetivo é encontrar grupos de entidades que sejam similares entre si em termos de sua composição total.

Enquanto o conceito de Modo Q é indubitavelmente útil em problemas puramente petrográficos, como quando tratando com composições minerais ou

suítes de rochas ígneas, seu uso em Exploração Geoquímica parece ser de aplicação mais limitada, os problemas sendo mais apropriadamente analisados por uma combinação de Análise de Componentes Principais e de Análise de Agrupamentos.

Os pontos fundamentais a serem tomados em consideração ao se comparar AF e ACP são, dentre outros, o de que em Análise Fatorial há vários procedimentos de estimativa, isto é, as estimativas não são únicas, em Análise de Componentes Principais há apenas um procedimento matemático, a solução é única se considerarmos **p** variáveis e **p** componentes sem rotação. Em Análise Fatorial a adição de outro fator pode modificar os anteriores, em Análise de Componentes Principais a adição de outra componente principal deixará as componentes principais anteriores inalteradas.

No estudo de bacias Eoproterozóicas-Neoproterozóicas do sul do Brasil estudou-se o Grupo Bom Jardim (o Grupo Bom Jardim na área é constituído da base para o topo pelas Formações Maricá (arenitos finos a muito finos com intercalações de pelitos e, subordinadamente, arenitos conglomeráticos e conglomerados), Hilário (afinidade alcalina, é dominada por andesitos, basaltos e dacitos, com arenitos conglomeráticos subordinados) e Arroio dos Nobres (composta de conglomerados polimíticos marrom claro com intercalações de arenitos conglomeráticos e, subordinadamente, psamitos e pelitos)), e sete componentes principais foram retidas para estudo, sendo que cinco variáveis hipotéticas representam 75,9% da variabilidade total presente, as duas últimas variáveis representam tão somente variações de elementos isolados. Isto significa um redução considerável na dimensionalidade do problema, pois em vez de analisarmos 42 variáveis passamos a lidar com cinco. As, Sb e Hg são utilizados como elementos farejadores para Au por sua associação com depósitos auríferos, terem concentrações mais elevadas, distribuições mais homogêneas e mais fácil determinação analítica. É necessário atentar, entretanto, para o fato de que nem todos os depósitos de Au têm associações anômalas com estes elementos, e que nem todas as zonas anômalas deles são associadas a depósitos auríferos. Em sedimentos de corrente estes elementos, juntamente com Ag e com metais básicos, têm a facilidade de poderem ser determinados com pequenas quantidades de fração fina dos materiais coletados. As parece ser o elemento que melhor resume as características apreciadas para funcionar como farejador para Au. Ag e Hg são farejadores de depósitos auríferos epitermais. Outros elementos comumente associados com mineralizações auríferas são Se, Te, Bi e Tl. Existem consideráveis problemas em se interpretar dados geoquímicos de sedimentos de corrente com base apenas nos resultados obtidos para Au, mesmo que se esteja prospectando ouro. Assim, é importante que se busque associações de elementos relacionados a algum evento ou a algum processo geológico atuante, pois há incontáveis relatos de casos em que os alvos se encontram em zonas de anomalias das associações e não simplesmente sobre anomalias

dos elementos principais envolvidos. A mineralização é resultado de um processo geológico que se impôs sobre a região com a interveniência de vários elementos conjuntamente, e não apenas resultado da atuação isolada do elemento considerado principal por seu valor econômico. Os sistemas hidrotermais relacionados com eventos intrusivos geralmente são mais focados e localizados nas anomalias geoquímicas que ocorrem ao redor dos corpos de minério, que são mais facilmente detectadas do que em bacias sedimentares, onde a quantidade relativa de metal quando comparada com a existente na fonte é pequena.

No estudo dos resultados do Grupo Bom Jardim se vê uma primeira componente responsável por cerca de 45,7% da variabilidade total, e representada por Ni, Co, Fe, Sr, V, Ca, Cr, Mg, Ti, Al, Na, Sc, Ga e Zr, e subordinadamente Zn, Mn, Ba e Li. Esta associação reflete muito a componente máfica/ultramáfica em muitos outros locais, segundo a bibliografia. Isto significa que os valores mais elevados destes elementos de forma conjunta delimitam zonas em que a contribuição máfica/ultramáfica se mostra mais efetiva, ou seja, nestas zonas os elementos mais importantes têm associação com este tipo de litologia. A primeira componente principal é a que responde pela maior parte da variabilidade presente nos dados, não significando ser a mais importante sob o aspecto prospectivo. Em alguns casos expressa tão somente uma composição litológica de maior influência, ou que imprime seu registro de modo mais proeminente na área.

A segunda componente principal (cerca de 12,7% da variabilidade total) está representada principalmente pelos elementos Sb, Tl, Cs, Rb e Be, e subordinadamente por As, K, Li e Sn. Esta associação, juntamente com vários dos elementos que compõem a anterior, também representa, em muitos outros locais da Terra, uma associação multielementar associada a depósitos do tipo epitermal. Em depósitos do tipo epitermal a ouro são freqüentes as associações de Au, Ag, Zn, Pb, Cu, Sb, As, Hg e Se nos de baixa sulfidização, e Cu, Au, Ag, As com Pb, Hg, Sb, Te, Sn, Mo e Bi nos de alta sulfidização. Como exemplos se pode citar McLaughlin, que é anômalo para Au, Ag, Sb, As, Hg e Tl (baixa sulfidização, que normalmente apresentam veios) e El Índio (anômalo para Au, Ag e Cu) e Rodalquilar, anômalo para Au, Cu, Te e Sn, e Paradise Peak, anômalo para Au, Ag e Hg, estes três últimos do tipo alta sulfidização (que são depósitos geralmente associados a rochas silicosas lixiviadas que se associam a fluidos gerados por ambientes vulcânicos com atividade hidrotermal).

A terceira componente principal no Grupo Bom Jardim (em torno de 8,2% da variabilidade total presente nos dados) é representada por zonas em que há associações de valores mais elevados conjuntamente de Ag, Au e Hg, com Bi subordinado; esta componente representa o comportamento de distribuição aurífera, ou seja, seus valores mais elevados mostram zonas em

que as condições para acumulação de ouro são mais favoráveis, pois que todos os elementos que compõem esta associação se relacionam com mineralização aurífera. Pode ser denominada de componente ouro. Esta terceira componente principal reflete anomalias conjuntas dos elementos Ag, Au, Hg e Bi, e não coincide necessariamente com anomalias individuais destes elementos.

O que cabe ressaltar é que as três primeiras associações citadas, que respondem por cerca de 66,6% da variabilidade total observada, se associam com associações do tipo epitermal em outros locais. Os depósitos do tipo epitermal costumam gerar anomalias de Au, Ag, As e Sb e às vezes de Hg em locais coincidentes (nas mesmas amostras de campo), se formam a profundidades rasas e têm o Au como principal elemento de importância econômica.

Exemplos de interpretação de associações destacadas pela aplicação da AF e/ou da ACP por diversos autores são, por exemplo, relacionar Ni - Cr - Co - V - Mn - Ti - As como representando associações máficas presentes na área, Ti - V - Cu, representando os locais de desenvolvimento de lateritas de certa região, Co, Ni e S representando uma combinação de processos sedimentares e de precipitação/sulfidização, indicando uma assembléia mineralógica de um depósito sulfetado. Em rochas graníticas a associação das variáveis Al, Fe, Mg, Ca e Ti determinadas pelo processo de diferenciação magmática, e Na e K representando metassomatismo e alteração pós-magmática, em granodioritos a combinação dos elementos Na / K / Ca sugerindo que nestas rochas seus teores foram determinados por diferenciação magmática em vez de por alteração pós-magmática. A associação Ti / V / Ni / (-SiO2) representando as rochas vulcânicas básicas presentes em certa área, Cr / Ni / Co em rochas ultramáficas e refletindo uma ocorrência de serpentinito previamente conhecida na área. Em outros exemplos a associação Cu / Cr / Ni, com alguma contribuição de Zn e Co, está associada a anfibolitos e Mn e Co indicam atividade de coprecipitação dos óxidos de manganês, com escores altos sobre xistos pegmatitizados. A associação Ni / Cr / V / Cu / Mg / Zn refletindo essencialmente um controle litológico máfico, Sr / Ba / P /Mg representando uma associação de elementos alcalino-terrosos com controle litológico por rochas intermediárias a plagioclásio a máficas, e Li / Pb / Sn / Zr refletindo associação típica de rochas félsicas . A associação SiO2 / TiO2 / Al2O3 / FeO / MgO / CaO / P2O5 / Zr / Sr / Ba / Li / Ni / Zn foi interpretada como representativa de um processo de diferenciação que gerou os líquidos a partir dos quais cristalizaram granitóides leucocráticos. A associação Ni / V / Cu / Cr / Sc / Co / Fe / Mg / Mn foi explicada como representando os metamorfitos do Grupo Açungui, em outro exemplo da bibliografia.

É indicado que se execute mapas com os escores obtidos, eles representarão o processo geológico associado à componente principal (ou ao

fator) retido, gera-se um mapa de processo geológico em vez de se gerar mapas de variáveis individuais, em muito maior número.

É usual se representar as primeiras componentes em gráficos que mostrem os percentuais representados pelas mesmas, em duas ou três dimensões.

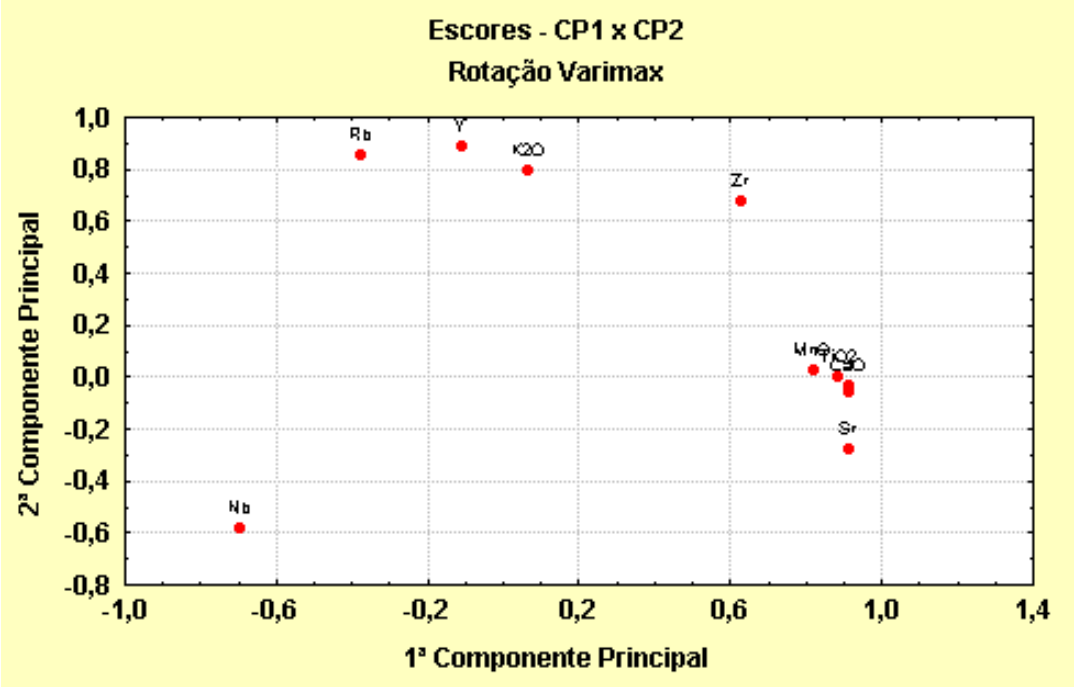


Figura 20 - Gráfico com as duas primeiras Componentes Principais

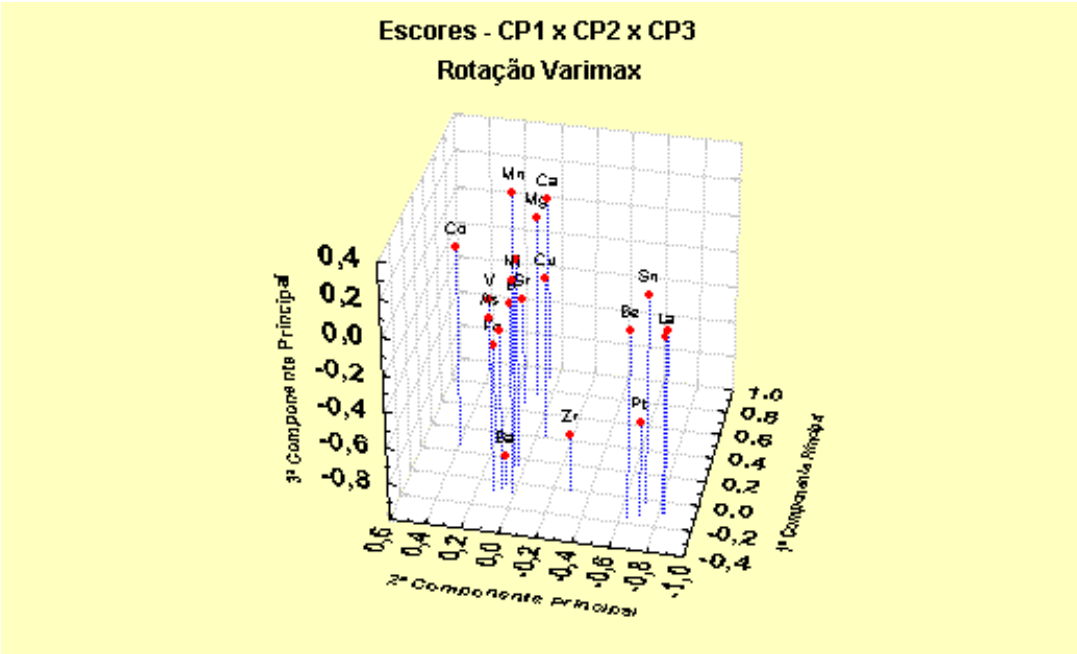


Figura 21 - Gráfico com as três primeiras Componentes Principais

A figura 22 mostra um exemplo de plotagem das cargas das variáveis (na vertical) com relação a cada fator (os fatores estão representados pela letra

F), e as figuras 23 e 24 mostram saídas de pesos fatoriais e escores fatoriais, respectivamente.

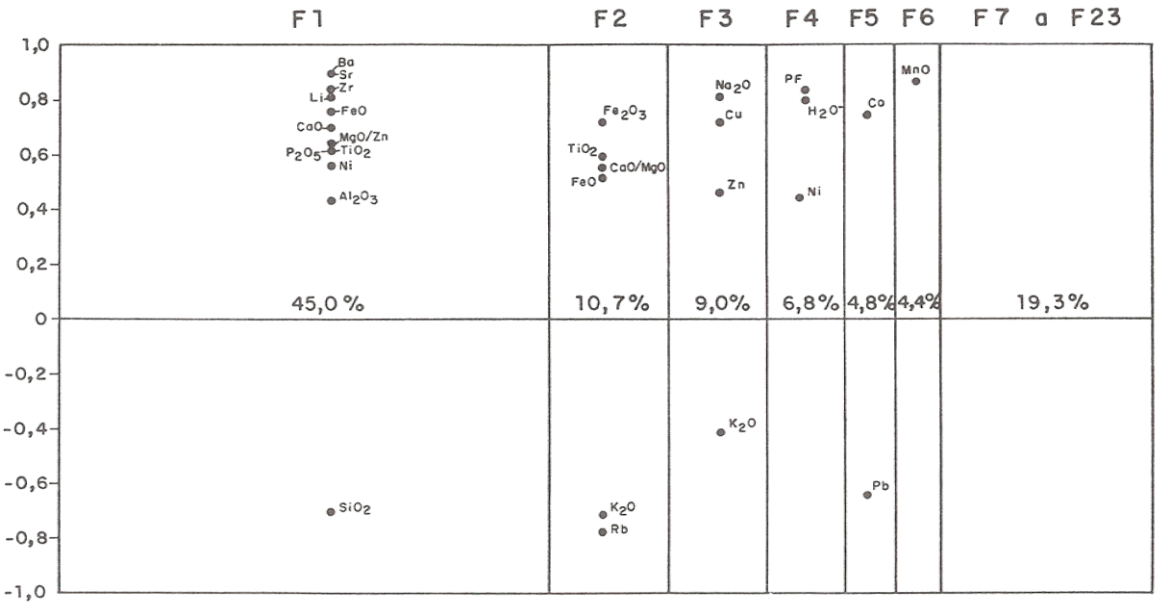


Figura 22 - Cargas das Variáveis x Fatores

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Moppm	0.05091	0.267901	0.246006	0.176305	0.404281	0.737690	0.040194
Cuppm	0.12943	-0.025487	0.256824	0.022450	-0.021122	0.843550	-0.095565
Pbppm	0.16035	0.518225	0.417571	0.207334	-0.006430	0.116372	0.182477
Znppm	0.67955	0.420480	0.072389	0.442242	-0.062961	0.107034	0.040094
Agppb	0.03904	0.027371	0.926889	0.039236	0.045063	0.269644	0.008190
Nippm	0.84596	0.214256	-0.039135	-0.249958	0.134045	0.129048	-0.001384
Coppm	0.85721	0.280854	0.055820	0.191734	-0.082176	0.059490	0.052365
Mnppm	0.62194	0.280765	0.090952	0.371556	-0.248608	0.096411	0.103088
Fe%	0.79032	0.189253	0.083285	0.496035	-0.052817	0.094958	0.009182
Asppm	0.27014	0.661572	0.245306	-0.099167	-0.034995	0.050657	0.173646
Uppm	0.17818	0.575319	0.088189	0.075544	0.727356	0.078222	0.020248
Auppb	-0.01628	0.019860	0.929974	0.012977	0.083025	0.120278	-0.020206
Thppm	-0.12880	0.040984	0.049711	-0.001702	0.824175	0.125570	0.197201
Srppm	0.83139	0.236455	0.093884	0.240199	0.088790	0.058135	-0.016620
Cdppm	-0.06572	-0.030886	-0.004834	0.063472	0.162940	-0.097105	0.771637
Sbppm	0.04703	0.704180	0.074464	-0.160324	-0.144833	0.174019	0.270934
Bippm	-0.00193	-0.077315	0.702816	-0.009294	-0.012359	-0.119124	-0.012427
Vppm	0.83372	0.151568	0.071931	0.394847	-0.078903	0.052066	0.035037
Ca%	0.87705	0.102635	0.011515	0.353043	-0.009744	0.013768	-0.059889
P%	0.45452	0.190181	0.001235	0.690741	0.014736	0.020401	-0.061220
Lappm	0.47754	0.510253	0.003327	0.528104	0.359132	0.014453	-0.022541
Crppm	0.86742	0.157817	-0.034454	-0.198590	0.108625	0.154968	0.025870
Mg%	0.93979	0.113181	0.029950	-0.062292	0.062313	0.012128	-0.082109
Bappm	0.54978	0.130357	0.039578	0.517350	-0.258903	0.265705	0.029843
Ti%	0.89698	0.087290	-0.023213	0.168924	0.115109	-0.040262	-0.057907
Al%	0.87085	0.295188	0.071200	0.300078	0.041007	0.010919	-0.039827

Figura 23 - Pesos Fatoriais

Case	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	0.40623	-0.32852	-0.42753	-1.36550	-0.19290	-0.80884	0.0039
2	-0.29852	-0.15352	-0.70473	-0.15480	0.00416	-0.01920	0.0248
3	-0.39972	0.51281	0.34575	1.83739	-0.66130	-1.27203	0.5320
4	-0.80877	-0.06955	-0.38860	-1.07128	-0.20171	-0.23643	-0.0386
5	-0.29269	-0.17476	-0.48688	-0.41489	-0.53831	0.40349	0.2222
6	-0.11483	-0.89554	-0.40748	-1.47497	-0.23972	0.18788	0.2145
7	-0.49071	-1.34243	-0.26416	-1.43016	0.16145	-0.58886	0.0378
8	0.81994	-1.15515	-0.27162	-0.72115	-0.31026	-0.30067	-0.0361
9	-0.45767	-1.01265	0.38089	-1.23942	0.95201	0.14778	0.2015
10	-0.16779	-0.65010	0.16560	-1.26174	-0.39859	0.20095	0.1347
11	1.02391	-0.09494	0.96391	-0.21160	-0.05016	-0.92667	0.4584
12	0.07594	0.30443	1.36397	-0.95208	-0.23215	-1.29187	0.2198
13	0.06466	0.86392	1.97815	-0.87627	0.28383	-2.65361	-0.3316
14	0.49635	-0.82476	-0.70411	-1.72294	-0.31297	0.00402	0.1559
15	-0.40113	0.20491	1.18250	-1.45783	0.17659	-0.04681	-0.1971
16	-0.14624	0.28304	1.30468	-1.18953	0.24177	0.39077	-0.1301
17	1.11258	1.35654	2.20471	-0.18677	0.75117	1.63218	-0.1421
18	0.15233	1.25692	1.01032	-0.89913	2.22699	1.14095	-0.0501
19	-1.00129	0.65955	7.70364	-0.06965	-0.02132	5.86835	-0.6691
20	0.05150	-0.74123	0.16359	-1.04427	-0.11811	0.34022	0.1832
21	-0.06830	0.14705	0.89205	-1.23086	0.33697	-0.92918	-0.0363
22	2.46142	-0.00204	1.27574	3.05639	-0.22453	4.23402	1.1630
23	1.49287	-0.86025	0.97989	-0.07628	-0.06328	-0.55427	0.4937
24	0.20479	0.36351	-0.33083	0.29205	-0.21372	0.90695	0.2852

Figura 24 - Escores Fatoriais

REFERÊNCIAS BIBLIOGRÁFICAS

Andriotti, J. L. S. – 2005 – Técnicas estatísticas aplicáveis a tratamento de informações oriundas de procedimentos laboratoriais. CPRM, 41 p.

Bolviken, B. A. - 1971 - A statistical approach to the problem of interpretation in geochemical exploration. Can. Inst. of Min. and Metallurgy, Spec. Vol. 11, p. 564 - 567

Bounessah, M. - August 2003 - The boxplot: a robust exploratory data analysis tool for the definition of the threshold for outlier data. Applied Geochemistry, Volume 18, Issue 8, Pages 1185-1195

Carranza, E. J. M. - 2009 - Geochemical anomaly and mineral prospectivity mapping in GIS / Handbook of exploration and environmental geochemistry, vol. 11, Elsevier, M. Hale editor - capítulo 3: Exploratory analysis of geochemical anomalies, p. 51 - 84

Cattell, R. B. - 1966 - The SCREE test for the number of factors. Mult. Behaviour Research, vol. 1, p. 245 – 276.

Dean, R. B. & Dixon, W. J. - 1951 - Simplified statistics for small numbers of observations. Anal. Chem. 23(1), p. 636 - 638

Doerffel, K. - 1962 - Beurteilung von analysenverfahren un-ergebnissen. Fresenius Z Anal. Chem., vol. 185, p. 1 - 98

Doerffel, K. - 1967 - Die statistische auswertung von analysenergebnissen. In: Schormüller I ed. Handbuch der lebensmittelchemie, vol. 2/2. Springer, Berlin, p. 1194 - 1246

Filzmoser, P., Garrett, G. G. and Reinmann, C.; – 2005 – Multivariate outlier detection in exploration geochemistry. Computers & geosciences, 31: p. 579 – 587

Hawkes, H. E. & Webb, J. S. – 1962 – Geochemistry in mineral exploration. Harper & Row, New York, 415 p.

Howarth, R. J. – 1983 – In: Sinclair, A. J. - Statistics and data analysis in geochemical prospecting. Elsevier, NY Cap. 3 – univariate analysis

Inácio, D.; Nogueira, P. & Noronha, F. – 2004 – Sugestão de procedimento base para o tratamento estatístico de dados referentes à prospecção geoquímica. Revista da Faculdade de Ciências, número 1, Edição Especial, p. 270 – 279.

Karger, M & Sandomirsky, S. – 2001 – Multidimensional statistical technique for detection of low contrast geochemical anomalies. Journal of Geochemical Exploration, 72, p. 47 – 58.

Lepeltier, C. – 1969 – A simplified statistical treatment of geochemical data by graphical representation. Economic Geology, vol. 64, p. 538 – 550.

Limpert, E.; Stahel, W. A. & Abbt, M. – May 2001 - Log-normal distributions across the sciences: keys and clues. BioScience, vol. 51, n. 5, p. 341 – 352

Matschullat, J.; Ottenstein, R. & Reiman, C. – 2000 (July) – Geochemical background – can we calculate it? Environmental Geology 39 (9), p. 990 – 1.000.

Miesch, A. T. – 1981 – Estimation of the geochemical threshold and its statistical significance. Journal of Geochemical Exploration, 16, p. 49 – 76

Parslow, G. R. - 1974 - Determination of background and threshold in exploration geochemistry. Journal of Geochemical Exploration, vol. 3, p. 319 - 336

Rantitsch, G. – 2004 – Geochemical exploration in a mountainous área by statistical modeling of polypopulational data distributions. Journal of Geochemical Exploration, 82, p. 79 – 95

Reinmann, C. & Filzmoser, P. – July 2000 – Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ. Geol., 39(9): p. 1001 – 1014.

Reinmann, C.; Filzmoser, P. and Garrett, R. G. – 2005 – Background and threshold: critical comparison of methods of determination. Science of the total Environment, 346: p. 1 – 16.

Rose, A. W.; Hawkes, H. E. & Webb. J. S. - 1979 - Geochemistry in mineral exploration. NY, Academic Press, 657 p.

Saager, R. & Sinclair, A. J. - 1974 - Factor analysis of stream sediment geochemical data from Mount Nansen Area, Yukon Territory, Canada. Min. Deposita, 9: p. 243 - 252

Sinclair, A. J. - 1974 - Selection of thresholds in geochemical data using probability graphs. J. Geochem. Expl., 3: 129 - 149

Sinclair, A. J. - 1976 - Applications of probability graphs in mineral exploration. Association of Exploration Geochemists, Rexdale, Ont., 95 p.

Sinclair, A. J. – 1986 – In: Reviews in Economic Geology, vol. 3, Society of Economic Geologists. Statistical interpretation of soil geochemical data, p. 97 – 115.

Sinclair, A. J. - 1991 - A fundamental approach to threshold estimation in exploration geochemistry: probability plots revisited. Journal of Geochemical Exploration, 41, p. 1 - 22

Sinding-Larsen, R. – 1977 – Comments on the statistical treatment of geochemical exploration data. Sciences de la Terre, série Informatique Géologique, n. 9, p. 73 – 90

Stanley, C. R. & Sinclair, A. J. - 1989 - Comparison of probability plots and the GAP statistic in the selection of thresholds for exploration geochemistry data. Journal of Geochemical Exploration, issues 1 - 3, p. 355 - 357

Tennat, C. B. & White, M. L. - 1959 - Study of the distribution of some geochemical data. Economic Geology, vol. 54, n. 7, p. 1281 - 1290

Tukey, J. W. - 1977 - Exploratory data analysis. Addison-Wesley, Reading, MA

Velasco, F. & Verma, S. P. – 1998 – Importance of skewness and kurtosis statistical tests for outlier detection and elimination in evaluation of geochemical reference materials. Mathematical Geology, vol. 30, n. 1, p. 109 – 128

Wellmer, F. -W. – 1998 – Statistical evaluations in exploration for mineral deposits. Springer – Verlag, 379 p.

Companhia de Pesquisa de Recursos Minerais

Sede
SGAN Quadra 603-Conjunto "J", Parte A, 1º andar
Cep: 70830-030 Brasília DF
Telefones: (61) 3223-1166; 3224-2069(PABX)
Fax: (61)3225-3985
E-mail: cprmsede@df.cprm.gov.br

Escritório Rio
Av. Pasteur, 404 Urca - Cep: 22290-240
Rio de Janeiro - RJ
Telefones:(21) 2295 5337 - (21)2295 0032 (PABX)
Fax: (021)2542 3647
E-Mail: cprm@rj.cprm.gov.br

Diretoria de Relações Institucionais e Desenvolvimento - DRI
Telefone: (21)2295 5837
Fax: (21)2295 5947

Departamento de Apoio Técnico - DEPAT
Telefone: (21)2295 5297
Fax: (21)2295 5947

Divisão de Documentação Técnica
Telefones: (21)2295 5997
Fax (21)2295 5897
E-Mail: seus@cprm.gov.br

Superintendência Regional de Belém
Av. Dr. Freitas, 3645 - Marco - Cep: 66095-110
Belém - PA
Telefones: (91)276 6976 - (91)276 8577 (PABX)
Fax: (91)276 4020
E-Mail: sureg@cprm-be.gov.br

Superintendência Regional de Belo Horizonte
Av. Brasil, 1731 - Funcionários - Cep: 30140-002
Belo Horizonte - MG
Telefones: (31)3261 3037 - (31)3261 0391 (PABX)
Fax: (31)3261 5585
E-Mail: suregbh@bh.cprm.gov.br

Superintendência Regional de Goiânia
Rua 148, 485 - Setor Marista - Cep: 74170-110
Goiânia - GO
Telefones: (62)281 1342 - (62)281 1522 (PABX)
Fax: (62)281 1709
E-Mail: cprmgo@terra.com.br

Superintendência Regional de Manaus
Av. André Araújo, 2160 - Aleixo - Cep: 69060-001
Manaus - AM
Telefones: (92)2126 0301 - (92)2126 0300(PABX)
Fax: (92)2126 0319
E-Mail: suregma@cprm-ma.gov.br

Superintendência Regional de Porto Alegre
Rua Banco da Província, 105 - Cep: 90840-030
Porto Alegre - RS
Telefones: (51)3233 4643 - (51)3233 7311 (PABX)
Fax: (51)3233 7772
E-Mail: sureg@pa.cprm.gov.br

Superintendência Regional de Recife
Av. Sul, 2291 Afogados - Cep: 50770-011
Recife - PE
Telefones: (81)3428 1511 - (81)3422 6448 (PABX)
Fax: (81)3447 4467
E-Mail: cprm@fisepe.pe.gov.br

Superintendência Regional de Salvador
Av. Ulisses Guimarães, 2862
Centro Administrativo da Bahia - Cep: 41213-000
Salvador - BA
Telefones: (71)230 0025 - (71)230 9977 (PABX)
Fax: (71)371 4005
E-Mail: suregsa@sa.cprm.gov.br

Superintendência Regional de São Paulo
Rua Costa, 55 Cerqueira Cezar - Cep: 01304-010
São Paulo - SP
Telefone: (11)3257 6430 - (11)3258 4744 (PABX)
Fax: (11)3256 8430
E-Mail: cprmsp@cprm.gov.br

Residência de Fortaleza
Av. Santos Dumont, 7700 - 4º andar - Papicu
Cep: 60150-163 - Fortaleza - CE
Telefones: (85)3246 1642 - (85)3246 1242 (PABX)
Fax: (85)3246 1686
E-Mail: refort@fo.cprm.gov.br

Residência de Porto Velho
Av. Lauro Sodré, 2561 - Bairro Tanques
Cep: 78904-300 - Porto Velho - RO
Telefones: (69)223 3165 - (69)223 3544 (PABX)
Fax: (69)2215435
E-Mail: secretaria@pv.cprm.gov.br

Residência de Teresina
Rua Goiás, 321 - Sul - CEP: 64001-570
Teresina - PI
Telefones: (86)222 6963 - (86)222 4153 (PABX)
Fax: (86)222 6651
E-Mail: cprm@te.cprm.gov.br

CPRM
Serviço Geológico do Brasil
Secretaria de Geologia,
Mineração e Transformação Mineral
Ministério de
Minas e Energia