TRATAMENTO ESTATÍSTICO DE DADOS GEOQUÍMICOS

UMA REVISÃO

JOSÉ LEONARDO SILVA ANDRIOTTI

MARÇO / 2022

Neste documento apresento uma resenha evolutiva de diversos artigos que tratam do assunto tratamento estatístico de dados geoquímicos, sem inserção de comentários pessoais sobre o assunto, de que já tratei em diversos artigos de minha autoria. Não cito nenhum destes que gerei, nos quais já emito opinião sobre a aplicabilidade e utilização nos projetos que a CPRM desenvolve, bem como isto ocorreu em cursos que ministrei na Empresa (e fora da Empresa) em diversas ocasiões.

A ideia do presente documento é, repito, apresentar como evoluiu este assunto segundo diversos autores (há muitos outros, apresento aqui apenas alguns), visando posicionar possíveis interessados no tema para que, de posse destas informações aqui disponibilizadas, possam já partir para estudos mais aprofundados sem necessitar fazer este levantamento histórico-evolutivo. O assunto continua a suscitar algumas dúvidas e mesmo aplicações diferentes na CPRM, esta diversidade é útil e faz parte do processo científico de aprendizado, mas devemos ter o cuidado de procurar, mesmo que respeitando os avanços constantes que fazem parte de qualquer procedimento científico, manter um mínimo de uniformidade nos produtos gerados.

Um ponto que desejo ressaltar é que o tratamento estatístico é uma ferramenta importante na interpretação dos dados disponíveis, mas o ponto mais importante de todos é a associação, a ligação com os processos geológicos atuantes na área de estudo. Buscamos, com a aplicação da prospecção geoquímica em determinada área de estudo, uma compreensão dos processos atuantes na mesma, e as respostas estão na geologia, buscamos o conhecimento da ocorrência dos recursos minerais porventura existentes em determinada área, e é a geologia da área que regula estas concentrações que nos interessam.

Os artigos aqui apresentados o são com sua referência bibliográfica seguida de trechos (traduzidos e/ou vertidos para o português de forma resumida) que exprimem as ideias centrais dos mesmos, aí sim com a interferência do autor naquilo que ele considera que realmente resume o artigo apresentado.

A apresentação se dá em ordem cronológica, começando com os artigos mais antigos. Ao final apresento um resumo de artigo clássico sobre o tema, **Aplicações de Gráficos de Probabilidade em Exploração Mineral**.

Ahrens, L. H. – 1954 – The lognormal distribution of the elements (2). Geochimica et Cosmochimica Acta, vol. 6, p. 121 – 131.

- É geralmente aceito que intervalos de classe menores devam ser usados para testar ajuste de distribuição com maior precisão, o que só é possível com número razoável de dados.

- Um exame da relação entre a razão da média aritmética para a média geométrica e a magnitude da dispersão é vista como razoavelmente quantitativa. A concordância é boa, exceto para dispersões extremamente altas.

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

- A distribuição lognormal é o caso, por exemplo, da distribuição de elementos traços em rochas para áreas em diferentes partes do mundo, para o tamanho de grãos em amostras de rochas sedimentares, e outros, de acordo com Coulomb (1959) e Cousins (1956).

- Ajuste à distribuição lognormal é geralmente o caso quando se trabalha com amostras de solo, e em reconhecimento de drenagem encontramos elementos-traço lognormalmente distribuídos em sedimentos de corrente.

- Para objetivos práticos se trabalha sobre curvas de frequências acumuladas, e sua construção segue os seguintes passos:

    - seleção de uma população tão grande e homogênea quanto possível
    - agrupamento dos valores em número adequado de classes
    - plotar o histograma
    - alisar o histograma para obter a curva de frequência
    - ao se substituir a escala aritmética por uma escala probabilística a curva acumulada é representada por uma ou mais linhas retas

- As frequências acumuladas são plotadas contra os limites inferiores das classes, ao se usar o centro das classes se introduz um erro nos parâmetros de tendência central (*background* e *threshold*) mas não na dispersão. É traçada a curva de limites de confiabilidade da curva, referida como sendo usualmente de 5%, a largura do intervalo de confiança é inversamente proporcional à importância da população considerada, maior a população mais estreito o intervalo referido, e se usa o teste de Pearson

para testar a lognormalidade. O autor cita que é mais fácil plotar e ajustar curvas cumulativas do que curvas de frequência usuais, pois é melhor (mais fácil) ajustar retas do que curvas sinoidais.

- No caso de *background*, uma linha reta mostra uma população única, no caso de uma curva de distribuição perfeita o valor do *background* corresponde à moda e mediana, seria a média geométrica dos resultados, que é mais estável que a média aritmética.

- Uma distribuição lognormal é completamente definida por dois parâmetros, a média geométrica e o coeficiente de desvio (*coefficient of deviation*). O coeficiente de desvio é um índice de dispersão específico para a distribuição de certo elemento em certo ambiente e expressa o grau de homogeneidade da distribuição. É uma característica muito importante da distribuição de um elemento em dada vizinhança, provavelmente relacionada com o tipo de dispersão geoquímica, mecânica ou química, e em consequência pode dar indicações sobre o tipo de anomalia encontrada, se singenética ou epigenética. Valores mais altos parecem indicar dispersão preponderantemente mecânica, mas isto não foi provado. Este valor varia entre -1 e 1.

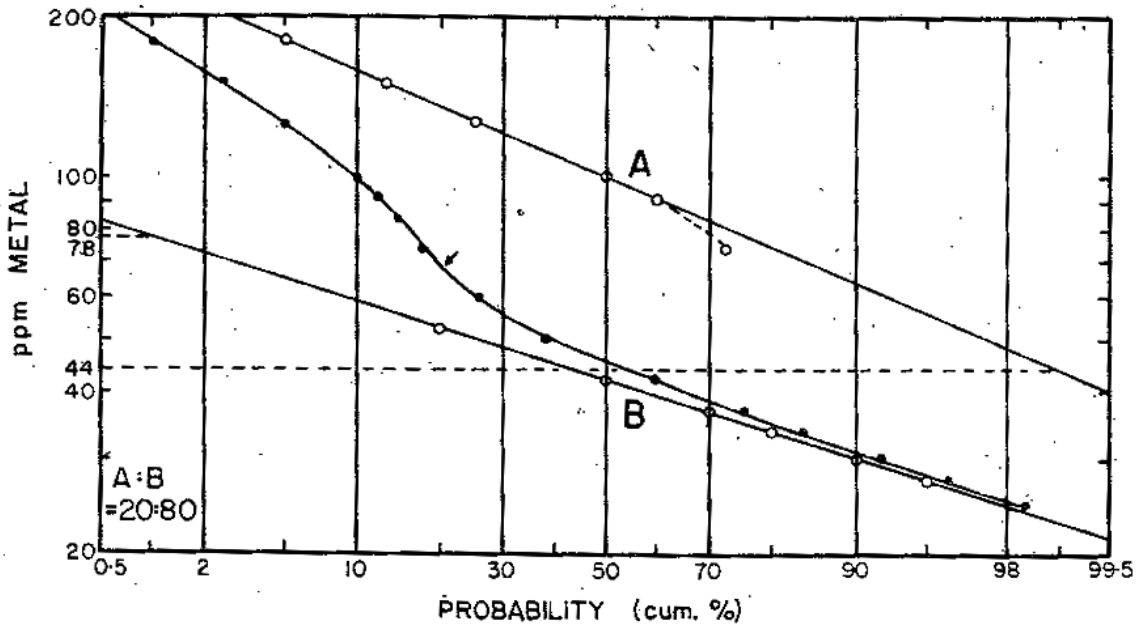
Sinclair, A. J. - 1974 - Selection of thresholds in geochemical data using probability graphs. J. Geochem. Expl., 3: 129 – 149.

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

- O método de média e desvios-padrão ignora o fato de que não existe nenhuma razão prévia para que se considere os 2,5% dados maiores como sendo anômalos, além do mais este método não considera que as populações anômala e de *background* têm em muitos casos superposição. Uma outra abordagem às vezes utilizada é se definir *thresholds* em pontos de máxima curvatura nos gráficos de probabilidades acumuladas.

- Nos gráficos de probabilidade que o autor apresenta os dados são acumulados do maior para o menor valor, e as vantagens citadas são a possibilidade de se examinar a forma da distribuição dos dados, a possibilidade de se estimar os parâmetros de populações normais e lognormais de forma rápida e acurada, a possibilidade de representar vários conjuntos de dados no mesmo gráfico e compará-los visualmente, e suas limitações seriam que os dados podem ter distribuições outras que não a normal ou lognormal, a exigência de pelo menos cem valores para sua construção e o fato de que o espalhamento dos dados pode ser grande demais para permitir uma análise confiável dos dados.

- Partição se refere aos métodos utilizados para extrair populações individuais a partir de distribuições polimodais. A figura que segue mostra um exemplo hipotético com um ponto de inflexão no percentil vinte, mostrando a presença de 20% de uma população A superior e 80% de uma população B inferior em termos de valores. O ponto superior de 180 ppm representa 1% do total dos dados, ou seja, representa (1 / 20 x 100) = 5 de percentual acumulado da população A porque neste extremo a população B não tem contribuição efetiva, consequentemente um ponto na população A é definido ao percentual acumulado igual a cinco no nível 180. Do mesmo modo o ponto no nível 150 representa (2,6 / 20 x 100) = 13% acumulado de A e um segundo ponto representativo da população A é obtido, este procedimento é repetido até que pontos suficientes são obtidos para representar e definir uma população A. Quando esta quantidade de pontos é alcançada uma linha é traçada através deles como uma estimativa da população A, e se faz o mesmo para a população B. Os pontos para as populações A e B estão representados como círculos abertos na figura. Na prática várias tentativas devem ser feitas para se obter um bom ajuste da mistura ideal devido à dificuldade de definição dos pontos de inflexão de modo acurado. Partição de curvas polimodais se faz da mesma maneira. A média geométrica pode ser lida no percentil 50 e a amplitude dos valores pode ser determinada nos percentuais acumulados 16 e 84 incluindo cerca de 68% dos dados, a amplitude englobando dois desvios padrão é assimétrica ao redor da média geométrica.



- O método descrito reforça a idéia de que *background* e *threshold* são populações que se superpõem, e também que o procedimento proposto

não se restringe à separação entre estas duas populações, mas também permite particionar os dados em diferentes populações simplesmente.

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

- Somente duas escalas são usadas normalmente em geoquímica, a aritmética e a logarítmica, esta última se ajusta à maioria dos elementos-traço.

- A curva sigmoidal é produzida pela superposição de duas distribuições teóricas.

- Em uma distribuição teórica os pontos de inflexão apresentam erros em torno de 5 a 10%, como em geoquímica as anomalias geralmente não ultrapassam este percentual em quantidade o uso do ponto de inflexão não contribui com muito erro, este método é mais preciso e vantajoso em casos em que a distribuição anômala é menos óbvia em termos de separação do *background*.

- A polimodalidade é comum, e pode gerar vários pontos de inflexão.

- Existe uma imprecisão inerente à acumulação de valores, na escala logarítmica deve-se dar atenção ao problema do limite inferior de detecção, pode haver problemas na cauda inferior da distribuição.

- As distribuições de *background* e anomalia são extraídas facilmente assim: as curvas lineares seriam tangenciais às caudas sigmoidais e os pontos de inflexão marcam as posições onde a distribuição do *background* está a uma distância abaixo da sigmoidal igual à distância da distribuição anômala acima.

- Fontes potenciais de erro são a posição dos pontos de inflexão, a curvatura das caudas sigmoidais, o desvio da sigmoidal simples, precisão gráfica e analítica.

- O método somente é aplicável na suposição de bimodalidade da distribuição, ou onde polimodalidade é suficientemente pequena para ser ignorada.

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.

- 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 (1972) 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 (1969) usam regressão múltipla para estimar a concentração de *background* em determinado local de amostragem, se o valor obtido for significativamente maior que o valor estimado de *background* para certo elemento então o local é considerado anômalo.

- Em todos os métodos utilizados por diversos autores a definição de anomalia está ligada ao desvio de algum valor da maioria dos demais, que formam a população de *background*.

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

- O autor apresenta a estatística GAP, que é introduzida para situações em que o *threshold*, se presente, é representado por um *gap* sutil em relação ao *background*. O primeiro passo é encontrar uma transformação dos dados que aproxime a distribuição da normalidade, depois transformar valores de acordo com a magnitude e padronizar os valores pela subtração da média e sua divisão pelo desvio padrão, para eliminar os efeitos de escala e permitir o uso de algumas tabelas de valores críticos. As diferenças absolutas entre os sucessivos valores resultantes na forma ordenada pode ser chamada de *gaps* padronizados, os maiores tendem a ocorrer próximos das caudas da distribuição e quase nunca em torno da média. Multiplica-se o valor do *gap* padronizado pela frequência esperada para o centro do *gap* conforme determinado a partir de uma curva normal ajustada. Obtém-se os produtos dos *gaps* padronizados pelas frequências esperadas aos *gaps* ajustados, o valor máximo de *gap* ajustado é a estatística GAP.

- Para dados com assimetria positiva, cauda nos valores mais altos, se faz $z = \ln (x - \alpha)$, onde α é uma constante que deve ser definida por processo interativo. Para dados com assimetria negativa (cauda na direção dos valores baixos) se faz

z = ln (α - x) onde α é derivado por acréscimo dele a partir de um valor inicial levemente superior ao valor máximo de x até que o valor absoluto da assimetria alcance um mínimo.

- Há três pré-requisitos para o uso da estatística GAP: independência das amostras, ausência de *bias* analíticos e a população de *background* deve ser proximamente ajustada a uma distribuição lognormal a três parâmetros.

- O procedimento completo para seu uso (estatística GAP) seria:
    - colocar as n amostras em uma sequência aleatória, denominando os dados por x, transformá-los, se necessário, para que a assimetria se aproxime de zero, e denominar os dados transformados por z.
    - padronizar os dados transformados de modo que a média seja zero e a variância seja um, e denominá-los por y.
    - ordenar os dados transformados padronizados e computar o *gap* ajustado, G(i), para cada par de valores através da amplitude (ou apenas acima da mediana) pela fórmula

$$G(i) = Fm[y(i + 1) - y(i)] \text{ para } i = 1, 2, ..., n - 1$$

onde y(i) e (y + 1) são os valores sucessivos e $Fm = 0{,}3989 \exp(- m^2 / 2)$, ou Fm pode ser lido de tabela da curva normal. A estatística GAP é o maior valor de todos os (n - 1) valores de G(i).

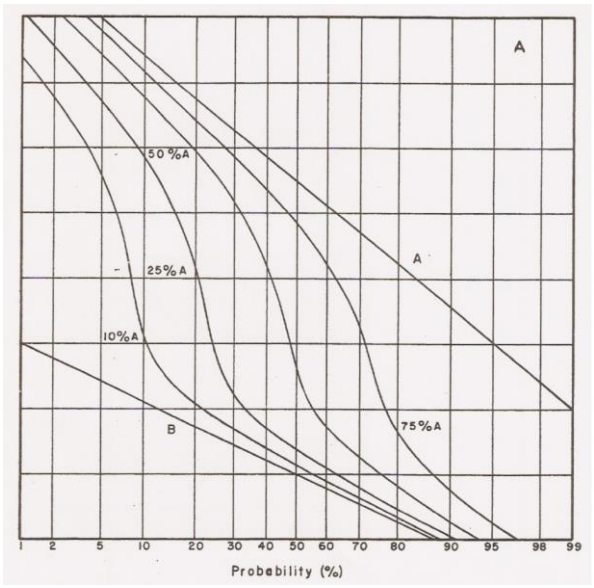
    - comparar a estatística GAP com os valores críticos para o número apropriado de amostras de acordo com tabelas específicas, ou, alternativamente, o maior valor *gap* ajustado serviria somente para identificar um *threshold* que possa ser testado estatisticamente em termos de sua significância por outros meios estatísticos.
    - a aplicação mais apropriada da estatística GAP é a separação de dados de duas ou mais populações que diferem apenas sutilmente, onde cada população é representada por diferentes mas não especificamente distintas modas no histograma.

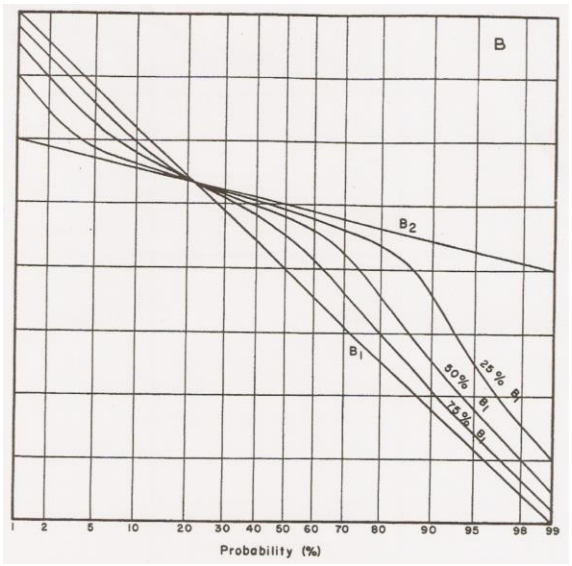
Howarth, R. J. – 1983 – In: Sinclair, A. J. - Statistics and data analysis in geochemical prospecting. Elsevier, NY Cap. 3 – Univariate Analysis.

- Os gráficos de probabilidade são muito sensitivos ao afastamento da normalidade e ao reconhecimento de múltiplas populações.

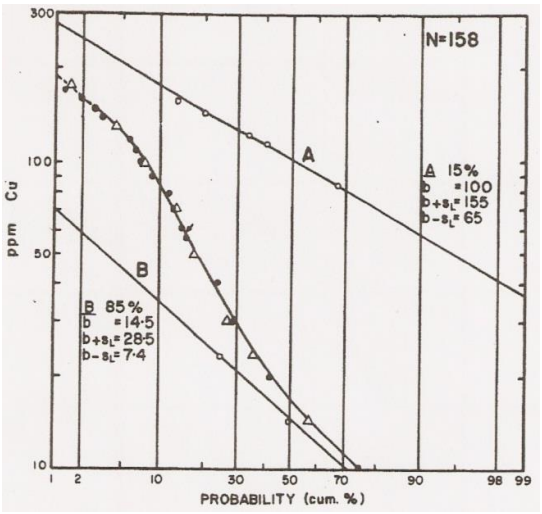
- O exame de muitos destes gráficos mostra que um ponto de inflexão ocorre na mistura em um percentual acumulado que coincide com as quantidades das duas populações presentes. Por exemplo, um ponto de inflexão que

aparece no percentil 15 indica 15% de uma população lognormal A (parte superior) e 85% da parte inferior de uma população lognormal B.





Duas populações hipotéticas combinadas em várias proporções relativas de acordo com o método gráfico descrito no texto. A - a bimodal mais comum encontrada em dados geoquímicos, as duas componentes se superpõem parcialmente, chamada de padrão *non-intersecting* por Sinclair (1976) para enfatizar o padrão de não intersecção das duas componentes sobre uma amplitude significativa de valores mostrada no gráfico de probabilidades. B - uma forma de intersecção relativamente incomum de gráfico de probabilidade para dados geoquímicos para os quais uma ou duas componentes têm uma amplitude (dispersão) relativamente pequena contida inteiramente na amplitude da outra componente.



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).

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

- Muitas variáveis das ciências da Terra incluindo elementos menores em solos mostram histogramas que se aproximam da forma da distribuição normal. Há também variáveis que são de uso rotineiro que mostram comportamento normal mas que incorporam na sua natureza uma logtransformação, como pH e tamanho de grãos na escala *phi*.

- É claro, nem todas as variáveis químicas em solos são lognormalmente distribuídas.

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

- Técnicas subjetivas de determinação de *threshold* baseadas em modelos, incluindo média mais dois desvios-padrão, são arbitrárias e ineficientes. Uma maior objetividade pode ser incorporada se for reconhecido que anomalia e *background* são respostas a processos dramaticamente diferentes, cada um caracterizado por sua própria função densidade de probabilidade (histograma). A determinação do *threshold* precisa ser vista como um procedimento de estimação no senso estatístico, sujeita a erros aleatórios e sistemáticos.

- Stanley e Sinclair (1989) propõem uma classificação de técnicas de seleção de *thresholds* em três categorias principais, descritas a seguir.

- Métodos experimentais - os que dependem da experiência dos técnicos e incluem o uso de tabulações de dados ou avaliação visual de histogramas, enfatizam as abundâncias absolutas de um modo altamente subjetivo. Assim, não são adequados para uso geral em que decisões precisam ser tomadas com base em comparações, como por exemplo contraste de anomalia, ou onde é crítica a importância de se classificar cada indivíduo como anômalo ou *background* (o que ocorre em trabalhos regionais).

- Abordagens baseadas em modelos subjetivos - usam algum tipo de modelo matemático ou estatístico formal para conjuntos de dados geoquímicos. Langford (1965) coloca que anomalia e *background* representam populações diferentes, geradas por diferentes causas e caracterizados por diferentes parâmetros estatísticos. Se estivermos com uma distribuição normal e utilizarmos a abordagem média mais dois desvios-padrão estamos assumindo que os 2,5% superiores são anômalos, claramente uma proporção arbitrária. Apesar das limitações, onde não houver evidência óbvia da presença de processos mineralizantes nos dados, e uma única população aparece, o uso de média mais dois desvios-padrão como *threshold* pode ser um fator de segurança útil para isolar alguns poucos valores para posterior avaliação.

- Abordagens baseadas em modelos objetivos - diferem das anteriores apenas no sentido de que os *thresholds* são definidos com base nos dados mesmo em vez de sobre decisões arbitrárias dos técnicos de exploração. Dois procedimentos nesta categoria são a abordagem dos *probability plots* (Sinclair, 1974 e 1976) e a estatística *gap* (Miesch, 1981). Esta última não foi muito usada, talvez por sua pouca aplicabilidade manual, pois que lançada em época de uso menos intensivo de computadores. Na prática

esperamos que as funções densidade de probabilidade para elementos particulares sejam diferentes em anomalias e *background*, o problema está em reconhecer esta diferença com confiabilidade.

- O autor cita tamanho de anomalia, ou proporção de amostras anômalas, e apresenta a fórmula $P_a = N_a / (N_a + N_b)$, onde N_a e N_b representam as quantidades de amostras de anomalias e *background* respectivamente, e também a fórmula $A_a = P_a \times S$, onde A_a é a área de uma anomalia e S é a área total de trabalho; o autor cita, porém, que na prática há poucas situações onde esta relação pode ser aplicada com confiança.

- A forma da anomalia pode influenciar tanto a aplicação de novas malhas como a interpretação, deve-se saber se uma anomalia é isotrópica ou anisotrópica, por exemplo, e neste último caso qual a orientação preferencial.

- O contraste de uma anomalia é definido como sendo a razão entre os valores médios da anomalia e do *background*, pela fórmula $C = x_a / x_b$, ou $C = x_a / (x_b + 2S_b)$, onde x_a e x_b são respectivamente os valores médios das populações de anomalias e de *background*, e S_b é o desvio padrão da população de *background*.

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.

- Observações aberrantes foram identificadas e rejeitadas usando 4 testes estatísticos diferentes (teste do coeficiente de assimetria, teste do coeficiente de curtose, teste de Dixon e teste de Grubbs); então, o valor médio remanescente foi avaliado e tido como valor de “consenso”.

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.

- Quando se trata com dados geoquímicos ou ambientais regionais as distribuições normal e lognormal são a exceção e não a regra. Porque a distribuição dos dados é tão importante? Porque muitas aplicações estatísticas, incluindo cálculos de níveis de probabilidades, se baseiam na hipótese de as distribuições serem normais. Dados geoquímicos e

ambientais mostram dependência espacial, e dados espacialmente dependentes não são, em geral, normalmente distribuídos.

- Há várias possibilidades de se testar a normalidade univariada, os testes mais populares são o de Kolmogorov-Smirnov, o qui-quadrado e o de Shapiro-Wilk, este último geralmente preferível aos outros dois.

- Os valores altos geralmente pertencem à mesma população dos demais dados? Em muitos casos provavelmente não, em geoquímica podem indicar tipos de rochas não usuais na área ou mesmo um depósito mineral, em ambiental podem indicar uma fonte de poluição.

- O desvio absoluto da mediana (MAD – *median absolute deviation*) é robusto contra um alto número de *outliers*, pelo uso da mediana e do MAD os dados não precisam seguir qualquer modelo de distribuição. Podemos também calcular média e desvio-padrão para dados logtransformados e então retrotransformá-los, fazendo isto obteríamos valores mais realísticos para média e desvio padrão mesmo para dados assimétricos.

- Média e desvio-padrão, que são os melhores estimadores de locação e espalhamento dos dados que seguem uma distribuição normal, estão longe de serem ideais quando usados para dados ambientais ou geoquímicos regionais. A razão para a forte anisotropia que exibem com freqüência é causada pela presença de mais de uma população ou processo. Em muitos casos a mediana é a melhor medida de posição para os dados. A média geométrica pode ser uma alternativa aceitável, mas tem alguns perigos. Como medida de dispersão o MAD ou o *hinge-spread* (Tukey, 1977, *Exploratory Data Analysis*) deveriam ser usados em vez de o desvio-padrão, que é muito vulnerável à existência de *outliers*.

- *Outliers* não influenciam métodos robustos, e métodos não paramétricos não se baseiam em suposições de distribuição, por isso são preferíveis aos métodos clássicos.

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

- Sob o ponto de vista geoquímico o termo *background* é equivalente à ausência de uma anomalia. As concentrações de *background* não são necessariamente iguais a baixas concentrações, entretanto. Um *background* geoquímico é caracterizado por variabilidade regional e é uma função de tempo.

- Processos induzidos naturalmente e antropogenicamente não somente levam a um alargamento da amplitude de variação dos dados coletivos (desvios-padrão maiores) mas também a distribuições multimodais. Idealmente cada moda corresponde a um processo relevante com sua distribuição normal subjacente.

- A citação de valores simples de *background* geoquímico não é útil para a caracterização do *background* geogênico nem para a determinação de uma contaminação antropogênica, pois que valores simples não aportam informação sobre desvio natural. É possível, entretanto, definir limites superiores para o *background* com uma determinada confiabilidade estatística. Em geral pode ser feita uma distinção entre métodos geoquímicos e estatísticos.

- Às vezes se faz necessária a utilização de estatística não paramétrica, uma aproximação pode ser a de que cada processo envolto na geração dos dados seja ele mesmo produto de um conjunto de dados com distribuição mais ou menos normal, e que o número n de processos se superpõe largamente.

- A simples comunicação de um valor médio sem o seu desvio padrão é de pouca utilidade, e somente pode ser usado em comparação com métodos geoquímicos. Faz sentido, entretanto, para mostrar intervalos de concentração (por exemplo, o intervalo normal de uma amostra é definido como valendo a média ± dois desvios-padrão, isto significa que 95% das amostras vêm deste intervalo).

- O autor cita Lepeltier (1969), que sugeriu um método de avaliação gráfica de somas cumulativas em escala gráfica bilogarítmica. Somente a concentração média do elemento (o valor do Clarke) é exigido. As idéias de Lepeltier se baseiam em que os valores dos elementos-traço mostram distribuição lognormal. O autor comenta que se torna óbvio que este método não pode ser aplicado a pequenos conjuntos de dados uma vez que eles serão muito poucos para mostrar e descrever sem erro uma quebra de curva.

- Com boa aproximação com Lepeltier existe uma técnica que requer curvas dos elementos individuais para mostrar as freqüências cumulativas relativas linearmente, Bauer e Bor (1993 e 1995) e Bauer e outros (1992) definem a primeira quebra da inclinação na curva como o limite superior de qualquer *background*. Esta técnica não requer nenhuma suposição prévia sobre a função de distribuição dos dados.

- Também relacionada com Lepeltier a técnica denominada de *Normal Range of a Sample*, de Hindel & Fleige (1991) define o percentil 97,5 de amostras antropogenicamente não influenciadas como o limite superior do

background. Esta abordagem parece problemática pois que não existe nenhum critério que permita que se distinga *a priori* entre amostras contaminadas e não contaminadas. A definição do *background* como "o intervalo normal de uma amostra" requer a exigência de uma distribuição lognormal. A determinação do *background* por média ± dois desvios-padrão é então feita graficamente, plotando as freqüências relativas acumuladas em escala logarítmica e lendo o percentil 97,5 ou por cálculo com o desvio padrão dos dados logtransformados.

- O autor cita também a técnica de regressão, em que as concentrações medidas de um elemento são confrontadas com outros elementos por meio da regressão linear de modo que o *background* pode ser calculado para cada ponto que confirma com as condições da regressão, e amostras que ficam acima do intervalo de confiança de 95% são definidas como antropogenicamente influenciadas.

- Análise Modal – este método objetiva a separação de funções de distribuição multimodais em várias distribuições normais, cujas médias correspondem com os respectivos (relativos) valores modais da função de distribuição inicial (Carral e outros, 1995). O *background* é derivado como sendo o limite superior do intervalo de confiança de 95% da média de uma distribuição normal do menor valor médio.

- Teste de *outlier* 4σ – assumindo uma distribuição normal dos elementos traços em amostras, e usando um conjunto de dados apropriado (entre 10 e 1.000 dados) um teste deste tipo pode ser aplicado, havendo mais de dez valores um resultado maior que média mais 4σ pode ser interpretado como sendo um *outlier*. Ele requer a eliminação de *outliers* potenciais (como por exemplo a partir das curvas cumulativas de frequência) do conjunto de dados e o cálculo da média e do desvio padrão para os dados remanescentes. O intervalo média ± dois desvios-padrão será visto como o intervalo normal do *background*.

- Técnica 2σ interativa – comparável com a técnica do teste de *outlier*, mas mais radical e matematicamente menos robusta, busca definir o *background* aproximando um intervalo normal (Erhardt e outros, 1998). Calcula-se média e desvio padrão dos dados originais, todos os valores fora do intervalo média ± dois desvios-padrão são omitidos, este procedimento é repetido até que todos os valores remanescentes fiquem dentro deste intervalo, o valor de média ± dois desvios-padrão calculado a partir do coletivo resultante é considerado refletir o *background* geogênico. Esta técnica constrói uma distribuição normal aproximada ao redor do valor modal dos dados originais, o resultado não é necessariamente o mesmo do obtido pela análise modal.

- Em resumo, há duas exigências para qualquer método estatístico que objetive determinar um *background* geogênico, a robustez e uma ampla aplicabilidade.

- Faz pouco sentido usar o *background* de dados – que é importante para modelos globais – para responder questões regionais ou, ainda mais difícil, problemas locais.

- Considerando a necessidade de encontrar um teste estatístico simples e robusto, tanto a técnica de dois desvios-padrão e uma distribuição calculada representam realisticamente os dados, todas as técnicas levam a uma aproximação com valores médios e medianos.

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

- É proposta técnica para detecção de anomalias geoquímicas de baixo contraste, usando-se uma medida estatística de contraste da anomalia, o valor τ, depois a seleção de um *background* para a população, e por fim a redução da dimensionalidade do espaço. Mapas de anomalias de τ podem ser rigorosamente interpretados com base em inferências estatísticas.

- Um problema que tem sido verificado é que os contrastes entre anomalias e *background* têm se mostrado estatisticamente insignificantes.

- A quantidade τ é uma função generalizada da distância de Mahalanobis entre uma observação individual e o centróide de uma população de *background* no espaço; como tal, é quase independente da dimensionalidade. Há que se considerar que o termo *background* é um termo grosseiro e indefinido.

- Uma medida de anormalidade de uma observação x pode ser o valor de g:

$$g = (x - \bar{x}) / s_b$$

  onde x é o logaritmo da concentração de certo elemento, e $\bar{x}$ e s_b são estimativas da média e do desvio-padrão, respectivamente, de uma distribuição de *background*. A interpretação geométrica do contraste da anomalia, g, é a distância entre dois pontos que é medida em termos de desvios-padrão em uma escala de logaritmos de concentrações. Se forem obtidos sobre a amostra na área de *background* o valor de g segue a distribuição de Student com (n – 1) graus de liberdade. Para grandes valores de n a diferença em relação à distribuição normal é pequena.

- Os procedimentos que operam com g podem ser vistos como testes estatísticos da hipótese de que x pertença a uma população normalmente distribuída, e o critério $|g| \geq g^*$, onde g^* é um *threshold*, é o critério mais poderoso para testar a hipótese.

- O problema de detectar anomalias geoquímicas tem duas feições especiais, a primeira é que não sabemos *a priori* se existem anomalias, e a segunda é que se anomalias existirem nós não sabemos a posição exata dos limites entre elas e o *background*.

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.

- 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. Alguns atributos, como por exemplo estaturas de pessoas, se ajustam bem às duas distribuições (normal e lognormal)

- Qual a diferença entre as variabilidades delas? Uma diferença importante é que os efeitos podem ser aditivos ou multiplicativos, o que leva a uma ou a outra distribuição. Veja-se o caso de dois dados, embora não se configure em nenhum dos tipos de distribuição discutidos, mas no caso aditivo eles variam de 2 a 12, e no multiplicativo variam de 1 a 36 com distribuição altamente assimétrica; assim, não podemos descrever os dois tipos de distribuição do mesmo modo.

- Dentre as propriedades básicas da lognormal está a de que somente valores positivos podem estar presentes, e é assimétrica à esquerda. Dois parâmetros são necessários para especificar uma distribuição lognormal. Tradicionalmente a média μ e o desvio padrão σ ou a variância σ^2 do log (X) são usados. Entretanto há claras vantagens em se usar a retrotransformação, onde os valores são, em termos de x, dados medidos:

$$\alpha^* = e^{\alpha} \qquad \text{e} \qquad \sigma^* = e^{\sigma}$$

- X é distribuída de acordo com a lei lognormal com a mediana μ^* e desvio-padrão multiplicativo σ^*. A mediana desta distribuição lognormal vale
med (X) = $\alpha^* = e^{\alpha}$ desde que μ seja a mediana de log (X). O parâmetro σ^*, que chamaremos desvio-padrão multiplicativo, determina a forma da distribuição. Distribuições são comumente caracterizadas por μ e por σ. Em aplicações para as quais a distribuição lognormal descreve adequadamente

os dados os parâmetros melhores para tal são a mediana α^* e o parâmetro de forma σ^*.

- 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

$$[\mu^* / \sigma^*, \mu^* . \sigma^*] = \mu^{*x} / \sigma^* \text{ contém 68,3\%}$$

e

$$[\mu^* / (\sigma^*)^2, \mu^* . (\sigma^*)^2] = \mu^{*x} / (\sigma^*)^2 \text{ contém 95,5\%}$$

- A soma de diversas variáveis normais independentes também é uma variável aleatória normal, e o produto de quantidades lognormais também segue a distribuição lognormal, a mediana do produto é o produto das medianas dos seus fatores.

- 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 desvio-padrão empíricos dos logaritmos dos dados são calculados e então retrotransformados, estes estimadores são chamados (X barra)* - a média geométrica dos dados, e s*. Mais robustos mas menos eficientes estimativas podem ser obtidas a partir da mediana e dos quartis dos dados. Tomando a natureza da distribuição em conta, a probabilidade do intervalo correspondente de (X barra) ± s fica em 88,4% em vez de 68,3%.

- Há um número de razões pelas quais as pessoas preferem a distribuição normal à lognormal. A principal parece ser a simetria, um dos princípios básicos do nosso modo de pensar. Duas outras razões se devem à simplicidade, pois adição é mais simples que multiplicação, e outra é que a distribuição normal foi conhecida e aplicada muito antes da lognormal. Esta preferência leva a que distribuições assimétricas que produzem altos valores tenham estes valores considerados como *outliers*, rejeitados e interpretados sem eles, reduzindo a assimetria mas introduzindo *bias*.

- Claramente a Química e a Física têm prevalência de multiplicadores em suas propriedades, o que causa um contraste óbvio: as razões que governam as distribuições de frequência na natureza usualmente favorecem a distribuição lognormal, mas as pessoas preferem a distribuição normal. Para coeficientes de variação baixos as distribuições normal e lognormal se ajustam bem, para estes casos é natural escolher a que for julgada mais apropriada para casos relacionados que exibam variabilidade crescente, que corresponde à lei

que governa as razões da variabilidade. Isto na maioria é a distribuição lognormal.

Jin-Yong Lee, Jeong-Yong Cheonb, Kang-Kun Lee, Seok-Young Lee and Min-Hyo Lee – 2001 - Statistical evaluation of geochemical parameter distribution in a ground water system contaminated with petroleum hydrocarbons - Journal of Environmental Quality 30:1548-1563.

Estudo feito em água subterrânea para identificar os processos dominantes que governam a distribuição de hidrocarbonetos contaminantes e parâmetros hidrogeoquímicos. Foram feitos estudos uni e multivariados. O estudo foi realizado em área a cerca de 45 km a SE de Seul, utilizando dados de 87 poços, nos quais se analisou, nas amostras coletadas, oito íons por cromatografia iônica e oito elementos por ICP-MS, além de compostos orgânicos analisados por cromatografia de massa.

Testes de Shapiro-Wilk e *box plots* foram incluídos, o primeiro mostrou que os parâmetros de interesse se aproximavam da distribuição normal para p = 0,05 sem transformação lognormal. Para estudos multivariados os dados foram padronizados ao escore **z** para as 23 variáveis estudadas, destas apenas 21 foram utilizadas, e 9 fatores explicaram cerca de 90% da variabilidade total dos dados, com rotação varimax após a extração dos fatores. Também foi executada *cluster analysis*, não restando locações de classificação duvidosa. A análise multivariada por fatores permitiu identificar diversos processos geoquímicos que controlam a composição ou as condições redox da água subterrânea, e a análise de *cluster* revelou que a distribuição dos parâmetros geoquímicos estava altamente associada com as ações de bombeamento e tratamento. Krigagem foi aplicada aos escores fatoriais.

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.

- *Outliers* são denominados *"wild" values*; as duas "*hinges*", a mediana e os dois valores extremos são conhecidos como o resumo dos cinco números.
- As *whiskers* incluem todos os dados, os *outliers* são definidos utilizando 1,5 vezes para mais ou para menos das "*fences*" superior ou inferior, respectivamente.
- *Boxplot* tem as seguintes vantagens: a) dá uma rápida idéia sobre a distribuição dos dados e sua estatística, como locação central, assimetria, *outliers*, etc; b) é muito robusto ao definir o *cutoff* para *outliers* que podem afetar parâmetros estatísticos na análise clássica, como média, desvio

padrão, etc; c) nenhum modelo particular é assumido para ajuste dos dados, o que evita a necessidade de transformar os dados.

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

- O principal objetivo do trabalho é demonstrar que a forma da distribuição dos dados de uma etapa geoquímica regional pode ser explicada por uma mistura de funções densidade de probabilidade paramétricas. Outro dos objetivos é a quantificação do *background*. Foram tratadas 695 amostras multielementares de sedimentos de corrente analisadas na fração –80 *mesh*, e valores inferiores ao limite de detecção foram considerados como tendo o valor do referido limite inferior. Foi usado o teste de qui quadrado para decidir sobre a rejeição da hipótese nula de distribuição normal com um nível de significância de 0,05. Para padronizar o procedimento de classificação os valores que se desviaram mais do que dois desvios-padrão da respectiva média aritmética ou lognormal foram interpretados como derivados de uma população diferente. Se a hipótese nula não pode ser aceita o segundo passo é a decomposição das distribuições observadas em suas componentes usando o *software* Tripod, que é baseado em uma decomposição de misturas de distribuições não linear por mínimos quadrados.

- De acordo com os testes estatísticos nenhum dos elementos seguiu uma distribuição normal, e também se demonstrou que os elementos transformados para logaritmos naturais também não exibiram distribuição normal. Be e La e os elementos dos seus grupos foram modelados pela distribuição normal, e As, Co, Pb e Zn mostram comportamento lognormal para logaritmos naturais, e alguns elementos como Ba, Ce, Na, Ni e V mostram distribuições normais e lognormais em diferentes grupos de bacias de captação.

- No conceito de Sinclair (1974 e 1991) as anomalias são caracterizadas por uma distribuição de dados que se sobrepõem com a população de *background*, mas as amostras anômalas estudadas não formam tais distribuições. Ainda de acordo com Sinclair uma aproximação de dois desvios-padrão da população de *background* foi usada para definir um *threshold* entre as amostras de *background* e as anômalas.

- O autor cita que Reinmann e Filzmoser (2000) e Reinmann e colaboradores (2002) sugerem que as técnicas estatísticas multivariadas não podem ser usadas para estudar dados geoquímicos polipopulacionais, e que é óbvio que *outliers* exercem severa influência sobre os resultados da PCA.

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.

- Se determinada área tiver rochas calcárias e ultramáficas e estivermos estudando Cr, cujas médias são respectivamente 2.000 ppm e 5 ppm, se não estudarmos os dados separadamente as eventuais mineralizações nas rochas calcárias dificilmente serão detectadas, e os descritores estatísticos terão seus significados reduzidos.

- *Outliers* podem ser essenciais para a determinação de eventuais mineralizações.

- Para determiná-lo se recomenda a utilização dos quartis e da mediana, que são robustos. Um modo simples de determinar os *outliers* pode ser

$$Q_1 - a(Q_3 - Q_1) < x < Q_3 + a(Q_3 - Q_1)$$

  onde Q_1 e Q_3 são o primeiro e terceiro quartis respectivamente, e **a** o fator de amplitude do intervalo.

- Alguns métodos estatísticos partem do pressuposto que determinada variável apresenta uma distribuição normal, o que em geoquímica raramente acontece. Zhang & Selinus (1998) salientam que as distribuições são geralmente assimétricas positivas, e que as distribuiçoes lognormais são apenas um caso específico das primeiras. Vistelius refere ainda que muitas das distribuições lognormais resultam da combinação de várias distribuições normais vizinhas.

- Um possível procedimento consiste na comparação da curtose, da assimetria e do teste de Shapiro-Wilks para os dados em bruto e seus logaritmos. Se uma distribuição for normal perfeita sua assimetria e curtose vão ser zero e o teste de Shapiro-Wilks será próximo de um. Há outros métodos, como o de *Box Plot*, que apresentam melhor resultado que a transformação logarítmica.

- O fato de um determinado elemento apresentar muitos *outliers* pode, por si só, indicar a existência de mineralizações.

- Como regra geral as fórmulas para determinar o intervalo de cada classe dependem da amplitude, se *outliers* não forem eliminados os intervalos se

tornam maiores e se perde detalhes e, eventualmente, informação importante.

- O coeficiente de correlação de Pearson é adimensional. Considerando que a correlação é fortemente influenciada por distribuições não normais e *outliers* se torna imperioso verificar o tipo de distribuição, se a variável não for normal deve-se proceder à sua normalização antes de se realizar a análise bivariada.

- Cluster é baseada no coeficiente de correlação ou no de distância, no primeiro se agrupa os elementos por comportamento semelhante, no segundo por quantidade semelhante. Para se usar o de distância os elementos precisam ter as mesmas unidades. A utilização de médias vai introduzindo distorções nos dendrogramas, que aumentam à medida que os grupos se tornam mais numerosos. Talvez a principal vantagem dos dendrogramas seja a possibilidade de detectar associações entre elementos, e consequentemente entre minerais.

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.

- Não há uma boa razão para continuar usando média $\pm$ 2 desvios padrão, originalmente proposto como "filtro" para identificar 2,5% dos dados no extremo superior da curva. Porque se tem 2,5% de *outliers* e não 5% ou 10%?

- Geoquímicos são tipicamente interessados em *outliers* como indicadores de processos geoquímicos raros, como mineralizações.

- Uma vez que *outliers* são definidos como sendo gerados por um outro processo, esta regra clássica não pode fornecer um *threshold* relevante. Em exploração geoquímica o intervalo entre média $\pm$ 2 desvios-padrão é definido com freqüência como sendo o "*background* geoquímico", reconhecendo que o *background* é um intervalo e não um simples valor.

- Valores baixos também podem ser importantes, eles podem indicar zonas de alteração (depleção de certos elementos) relacionados a acumulações minerais próximas (ocorrências). A implicação é que esta regra clássica não é válida.

- Para selecionar melhores métodos para tratamento de dados geoquímicos as propriedades básicas dos conjuntos de dados geoquímicos necessitam ser identificadas e entendidas, elas incluem:

- Os dados são espacialmente dependentes
- Em cada local amostrado vários processos têm influência no valor analítico determinado

- Para muitos testes estatísticos é necessário que as amostras provenham da mesma distribuição – isto não é possível se diferentes processos influenciam diferentes amostras

- A média poderia ser substituída pela mediana e o desvio-padrão pelo MAD. Estes estimadores são robustos contra valores extremos. Outra solução 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 estendida 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.

- Adicionamos a abordagem mediana $\pm$ 2 MAD porque, em adição ao *box-plot*, é a mais inteligível abordagem robusta, é uma analogia direta a média $\pm$ 2 desvios-padrão.

- Geralmente o método mediana $\pm$ 2 MAD dá o mais baixo *threshold* e, assim, identifica o maior número de *outliers*, seguido pelo *boxplot*. O *threshold* definido pelo *boxplot*, o "*inner fence*", é em muitos casos próximo, mas menor, do que o *threshold* obtido a partir de dados logtransformados usando a regra média $\pm$ 2 desvios-padrão. Então, mediana $\pm$ 2 MAD sempre resulta no *threshold* mais baixo, o *boxplot* em segundo e a regra clássica em terceiro, ou seja, o mais alto *threshold*. Como os dados geoquímicos geralmente são assimétricos à direita e muitas vezes se assemelham a dados lognormais, a segunda simulação fornece a explicação para o comportamento observado.

- Verdadeiros *outliers*, em vez de valores extremos, são derivados de processos diferentes e não de distribuições normais.

- Resultados das simulações sugerem que o *boxplot* é melhor se a percentagem de *outliers* nos dados estiver entre 0 e 10%, no máximo 15%; acima deste valor a abordagem mediana $\pm$ 2 MAD pode ser aplicada. O método clássico somente se mostra bom se não existirem *outliers.*

- Uma diferença importante entre a *boxplot* e os outros dois métodos é o fato de que as "*fences*" (os limites dos *outliers*) da *boxplot* não são necessariamente simétricos ao redor do centro (mediana). Eles são simétricos somente se a mediana for exatamente o meio entre as "*hinges*", que são os quartis superior e inferior. A diferença é mais realística para dados geoquímicos que a hipótese de simetria.

- Na maioria dos casos as distribuições geoquímicas para elementos menores e traços são mais próximas da distribuição lognormal do que da normal.

- Nos *plots* Q-Q ou Q-*normal plots* (quantis da distribuição dos dados são plotados contra os quantis de uma distribuição normal 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*.

- É recomendável que, ao plotar os dados, os *outliers* tenham um símbolo próprio que os identifique.

- Como conclusões, dos três procedimentos o *boxplot* é o mais informativo se o número de *outliers* verdadeiros nos dados estiver abaixo de 10%.

- O uso continuado da regra média $\pm$ 2 desvios-padrão é baseada em um mau entendimento, os geoquímicos desejam identificar *outliers* e não os extremos de uma distribuição normal ou lognormal, que é o interesse dos estatísticos.

- A inspeção gráfica da distribuição por meio da probabilidade acumulativa (ou Q-Q *plot*) antes de definir os intervalos de *background* ou *threshold* é uma necessidade absoluta. Somente o Q-Q *plot* combinado com a representação espacial pode fornecer uma resposta mais clara para as questões de *background* e *threshold*.

Um esquema proposto para a inspeção dos dados e seleção dos limites de variação do *background* está a seguir:

- 1 - Plotar as distribuições empíricas em escalas linear e de probabilidade normal, e os *boxplot* de Tukey. Se houver poucos dados extremamente separados da massa de dados prepare um conjunto com estes dados separados do todo.

- 2 – Computar S (desvio padrão) dos dados e do subconjunto gerado, se for o caso. O coeficiente de variação (CV) é um guia útil para a não normalidade, alternativamente a assimetria dos dados pode ser estimada.

- 3 - Se o CV for maior que 100% *plots* em escala logarítmica devem ser preparados. Se o CV estiver entre 70% e 100% a inspeção de *plots* em escala logarítmica será informativa. Outro guia útil é a razão valor máximo / valor mínimo, se exceder duas ordens de magnitude os *plots* logarítmicos serão informativos, se a razão estiver entre 1,5 e 2 ordens de magnitude os *plots* logarítmicos serão talvez informativos.

- 4 – Calcular as "*fences*" por mediana $\pm$ 2 MAD, se o passo 3 indicar que a lognormalidade for informativa, logtransformação dos dados, repetição dos cálculos e retrotransformação aos valores naturais.

- 5 – Preparar mapas usando os valores de "*fences*", plotando mediana e quartis, e ver se as concentrações de valores em certos intervalos podem ser associadas a diferentes processos.

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

- Em geoquímica *outliers* são geralmente considerados como observações resultantes de um processo secundário e não como valores extremos de uma distribuição de *background*.

- Quando os computadores ainda não eram usuais se usava a média com dois desvios padrão de afastamento dela para definir as anomalias.

- *Outliers* multivariados podem agora ser definidos simplesmente como as observações que tenham uma larga distância de Mahalanobis (quadrática). No caso multivariado um quantil de 98% da distribuição do qui quadrado poderia ser usado para definir seu limite. A distância de Mahalanobis precisa ser definida por método robusto para dar medidas confiáveis no reconhecimento de *outliers*, pois ela é muito sensitiva à presença de *outliers*.

- *Outliers* são pensados como sendo observações oriundas de uma ou mais distribuições diferentes, e extremos são valores afastados do centro mas que pertencem à mesma distribuição. Em um mapa os *clusters* de *outliers* indicariam algumas regiões com estrutura de dados completamente diferente das outras.

- O *plot* de qui quadrado é útil para visualizar desvios de dados a partir da normalidade nas caudas.

Grünfeld, K. – 2005 - Dealing with outliers and censored values in multi-element geochemical data – a visualization approach using XmdvTool Applied Geochemistry 20 (341–352).

Grünfeld, K. – 2005 – Visualization, integration and analysis of multi-element geochemical data. PhD Thesis, Stochholm, 54 p.

- Como uma regra geral quase todas as variáveis medidas em geoquímica regional não apresentam distribuição normal nem lognormal, no trabalho apresentado o foco é a exploração e remoção de dados censurados e de *outliers* de dados de estudos geoquímicos multielementares regionais. As decisões a serem feitas podem ser remoção ou substituição.

- Um limite de significância foi estabelecido anteriormente por outros autores como sendo menos de 1% dos dados, multiplicando o valor inferior por 0,33 a 0,50 do limite inferior de detecção.

- As técnicas estatísticas paramétricas padrão para dados multivariados, no caso de remoção de *outliers*, não são robustas e não podem ser aplicadas a dados geoquímicos brutos.

McQueen, K. G. - 2005 - Identifying geochemical anomalies – Internet.

Anomalias geoquímicas podem ser resultado de processo não usuais de concentração, ou por processos ativos durante longos períodos, ou contaminação artificial, ou erro ou ruído analítico. Anomalias negativas também podem ser importantes, como, por exemplo, para refletir depleção de alguns elementos durante a alteração da rocha hospedeira acompanhando a formação de minério.

- Métodos estatísticos univariados podem ajudar a identificar o tipo de distribuição dos dados, a presença de múltiplas populações e a presença de *outliers*. Box e Whisker *plots* são outra forma conveniente de se examinar a distribuição de freqüência de um conjunto de dados e para se comparar distribuições de freqüência de múltiplos conjuntos de dados. Eles mostram a mediana, primeiro e terceiro quartis, uma indicação definida como valendo 1,5 do comprimento do box (o IQR, *interquartile range*) em direção ao máximo e ao mínimo e setas que se estendem até os valores máximo e mínimo. A caixa central contém 50% dos dados. Valores fora dos limites dos *whiskers* são considerados *outliers* e valores maiores que três vezes o IQR (comprimento do *box*) além dos limites do *box* também são referidos como *outliers*.

- Hawkes and Webb (1962) consideravam que em uma população normal os valores fora do intervalo de dois desvios padrão eram anômalos. Segundo Reinmann e Filzmoser (2000) a maioria dos conjuntos de dados geoquímicos têm distribuições entre normal e lognormal.

- Um método alternativo é o de definir o threshold a dois MAD (*Median Absolute Deviations*) a partir da mediana. A mediana estará afastada da média nos dados assimétricos e os valores extremos terão menor influência. O MAD é definido como o valor mediano dos desvios absolutos a partir da mediana de todos os dados (Tukey, 1977). A abordagem MAD é melhor aplicada quando os dados contêm menos que 10% de outliers.

- 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.

- A forma e o tamanho de anomalias multivariadas podem ser quantificadas pela matriz de covariância usando uma medida chamada distância de Mahalanobis, formando uma elipse quando houver somente duas variáveis e um elipsóide quando houver mais do que duas variáveis. Os dados podem ser observados em termos de sua distância de Mahalanobis até o centróide da distribuição e os *outliers* podem ser definidos.

Mapas de contorno são fáceis de visualização mas podem não ser o melhor método de apresentação de muitos dados geoquímicos, pois muitos deles não possuem características próprias para contornos. Mapas com círculos com diâmetros variando em função do valor podem representar bem muitos valores.

Stanley, C. R. - February 2006 - Numerical transformation of geochemical data: 2. Stabilizing measurement error to facilitate data interpretation - Geochemistry: Exploration, Environment, Analysis, v. 6, no. 1, p. 79-96.

As distribuições de Poisson, Binomial e Hipergeométrica são tipicamente usadas para descrever os erros de amostragem. Enquanto que linhas retas (constante, proporcional ou afim) são usadas para descrever erros analíticos.

Stanley (2006) indicou que a transformação para o objetivo de criar dados normalmente distribuídos é provavelmente mal utilizada, pois vários conjuntos de dados geoquímicos são multimodais, e desse modo não podem ser “normalizados” com uma simples transformação monotônica que não altera

fundamentalmente a natureza dos dados. Stanley citou ainda que a aditividade é uma característica que ocorre comumente em concentrações geoquímicas de elementos-traço, de modo que uma transformação para criar aditividade é raramente necessária, a menos que concentrações de óxidos principais estejam sendo consideradas.

Transformação para estabilizar o erro de uma variável geoquímica cria uma nova variável que é homoscedástica (tem variância do erro constante ao longo de toda a faixa de variação da mesma).

Três distribuições estatísticas têm sido tradicionalmente empregadas para estimar a magnitude dos erros de amostragem, a Binomial, a Hipergeométrica e a de Poisson. Por exemplo, se considerarmos amostras contendo uma pequena quantidade de pepitas, típico de minérios de ouro, a distribuição de Poisson é apropriada para ser usada, pois tem uma variância que é funcionalmente relacionada com o número de eventos raros observados (as pepitas).

Se um elemento ocorre em um grão mineral que é comum em uma suíte de amostras o erro de amostragem correspondente não pode ser descrito por uma distribuição de Poisson, e se necessita de uma distribuição alternativa. Considerando que a concentração de um elemento em uma amostra é tecnicamente igual à probabilidade de selecionar aleatoriamente um átomo do elemento de interesse da amostra, se isto for tecnicamente possível, há efetivamente dois tipos de átomos na amostra: os do elemento de interesse (por exemplo, Cu) e os demais (não Cu). Coletar (amostrar) um grande número de átomos aleatoriamente para produzir uma “amostra”, dado que haja uma população suficientemente grande de átomos de modo que a seleção aleatória de átomos sem substituição não altere as probabilidades, ocorreriam erros que são distribuídos de modo Binomial. Note que Binomial significa dois nomes, no caso Cu e não Cu. Os elementos podem residir em mais de um tipo de grão mineral dentro de uma amostra (como calcopirita, bornita, calcocita) que podem diferir em tamanho e composição. Assim, os erros de amostragem não seriam verdadeiramente binomiais em distribuição, podem ter distribuição resultante de amostras que são misturas físicas de diversos tipos de grãos minerais de diferentes tamanhos e composições que contenham o elemento de interesse, e em que cada um dos tipos de grãos minerais pode ser binomialmente distribuído. Se o tamanho da população for pequeno relativamente ao tamanho da amostra coletada a probabilidade provavelmente mudará. E então uma distribuição alternativa precisa ser considerada. A distribuição que se aplica a estes casos envolvendo “amostragem sem reposição” é a distribuição Hipergeométrica.

A transformação da concentração dos dados em uma nova variável que exiba uma distribuição normal, próxima da normal ou simétrica geralmente não é possível em função do caráter multimodal de muitos conjuntos de dados geoquímicos. Como resultado, a transformação para obter este resultado é geralmente mal dirigida e com frequência não possível. O procedimento de análise de dados empregado (uma transformação) não é baseado na interpretabilidade dos resultados, sendo,

assim, científico. Como resultado, o uso da estabilização de máxima variância ou de variância do erro resultará em uma avaliação mais objetiva dos dados geoquímicos, e fornece novas oportunidades para reconhecimento de novas características nos dados que não tiverem sido previamente identificadas.

Grunsky, E. C. – 2007 - The evaluation of geochemical survey data: data analysis and statistical methods using geographic information systems - GIS for the earth sciences - Special Publication 44 – Chapter 12, p. 229 – 283.

Grunsky, E. C. - 2007 - The interpretation of regional geochemical survey data - advances in regional-scale geochemical methods. Paper 8 - In "Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration" edited by B. Milkereit, p. 139-182.

- A avaliação e interpretação de dados geoquímicos depende do entendimento da natureza do material amostrado. Por exemplo, a escolha da fração granulométrica pode ter grande influência na interpretação da geoquímica de uma área, diferentes frações podem refletir diferentes processos geológicos, com respostas geoquímicas muito diversas. Também é importante considerar que avaliar dados geoquímicos sem considerar controle analítico de qualidade pode ser perigoso.

- Visualização é um dos modos mais efetivos de se avaliar dados geoquímicos, o olho humano é de longe mais efetivo no reconhecimento de padrões nos dados do que tabelas de dados.

- Análise exploratória de dados se relaciona com o objetivo de detectar *trends* ou estruturas nos dados.

- *Background* geoquímico representa uma população que reflete superfície não mineralizada, e pode ser uma mistura de várias populações. A separação desta população em subgrupos similares que representam populações multivariadas normais é importante e forma a base das abordagens.

- Problemas que comumente ocorrem no processamento de dados incluem a presença de muitos elementos com distribuição “censurada”, ou seja, com valores inferiores ao limite de detecção sendo reportados apenas como “menor do que”, distribuições de dados não normais, falta de resultados para algumas variáveis em algumas amostras, e combinação de grupos de dados que mostram diferenças distintivas entre elementos, que podem ocorrer em função de diferentes limites de detecção, instrumentos ou pobre controle de qualidade de procedimentos, e também o problema da soma

constante para dados composicionais. Todos estes problemas causam dificuldades na interpretação de dados, especialmente no uso de técnicas estatísticas.

- Dados geoquímicos são reportados como proporções (% em peso, ppm, etc), para determinada observação a soma é 100%, o que resulta que algumas medidas “crescem” enquanto outras são “forçadas” a “decrescer” para manter a soma constante. Medidas estatísticas sobre dados com soma constante, como correlação e covariância, podem ser alteradas em função disto. O problema é raramente severo pelo fato de que a análise multivariada usa elementos-traço, e usualmente não representar a composição total, mas apenas um pequeno percentual da soma total. Entretanto é perigoso fazer a afirmação de que nenhum efeito ocorre nas medidas estatísticas nestes casos.

- A principal vantagem do *box plot* é que sua forma não depende da escolha dos intervalos, como nos histogramas.

- O coeficiente de variação é útil porque a dispersão é expressa como um percentual, de modo que pode ser usada como uma medida relativa para comparar diferentes elementos.

- Num estudo de caso apresentado, é relatado que quando a detecção de “anomalias” era menos óbvia foi usado o limite do percentil 98 para representarem, os valores superiores, os alvos.

- Um *outlier* pode ser definido como uma observação distintivamente diferente daquelas com que está estreitamente associada. Se um *threshold* foi definido, então um *outlier* o excede. Podem significar erros também.

- Métodos robustos de estimação são usados para minimizar os efeitos da presença de dados atípicos, e podem ser aplicados tanto a dados uni como multivariados. Dentre estes métodos se pode citar as *Trimmed Means*, estimativas robustas razoáveis podem ser obtidas fazendo-se corte dos dados em 5% ou 10%, isto é feito selecionando-se as observações que estiverem abaixo dos percentis 90 e 95 ou acima dos percentis 10 e 5, respectivamente. Outro método é o *Dominant Cluster Mode*, que computa uma média e um desvio-padrão iniciais, as observações que excedam K desvios-padrão são eliminadas e média e desvio-padrão são recalculados, de modo interativo até que não permaneçam mais observações que excedam K desvios-padrão. O método denominado *L-Estimates* se baseia em combinações lineares de algumas estatísticas, a Tri-média é um *L-Estimate* calculado como valendo 0,5 mediana + 0,25 *hinges*. O método *M-*

Estimates usa parâmetros baseados nas estimativas de máxima verossimilhança, o método "*bi-square*" define uma constante "c" (*threshold*), qualquer valor maior que "c" é atribuído ao valor zero; o método Huber pondera *outliers* baseado na constante "c", onde *outliers* são considerados os valores que superem um determinado número de vezes o valor "c". Outro método utilizado é o de *Huber W-Estimates*, em que estimativas preliminares de média e desvio-padrão são baseados na média (M) e no IQR dividindo-se por 1,35 (S* - IQR/1,35). A partir destas estimativas iniciais os resíduos são computados baseados na fórmula $(x_i – M) / S^*$, os resíduos que excedem este valor são eliminados, em processo interativo que continua até que os resíduos não mais excedam o valor de *cut-off*.

- Muitos procedimentos estatísticos assumem que as populações sob teste sejam normais, se *outliers* estiverem presentes a premissa de normalidade é violada. A aplicação de transformações nos dados pode mascarar a presença de múltiplas populações e *outliers*. Se transformações forem aplicadas aos dados originais para minimizar a assimetria modificações ocorrem transformações que podem ser aplicadas são a *linear scaling*, com y = k.x ou y = (x_i – média) / desvio-padrão, outra tranrformação é a exponencial, em que $y = e^x$, outra é a *Power* generalizada *Box-Cox*, em que $y = x^{\lambda} - 1) / \lambda$, y = ln (x) para λ = zero. As transformações do tipo *linear scaling* não modificam a forma das distribuições mas as medidas de dispersão (variância) podem mudar. As transformações logarítmica, exponencial e *Box-Cox* generalizada modificam tanto a forma como a característica da dispersão, e são as mais comumente usadas.

- Às vezes se torna necessário fazer uma padronização dos dados de diferentes origens (*levelling*), o primeiro passo é determinar qual dos conjuntos de dados será o padrão, o que dependerá de diversos fatores, como proximidade espacial dos conjuntos de dados, acuracidade e precisão do conjunto padrão, e se o conjunto padrão tem suficientes amostras e elementos de modo que os demais conjuntos possam ser nivelados a ele.

- Muitos métodos multivariados requerem estimativas de correlação ou covariância de modo que as interrelações entre as variáveis podem ser quantificadas. A influência dos *outliers* pode ser reduzida pela aplicação de métodos robustos para a estimação das médias, correlações e covariâncias entre as variáveis. Na estatística multivariada a distância de uma determinada observação a um centroide é estimada pela distância de Mahalanobis (equivalente multivariado do desvio-padrão normal), que depende de uma estimativa da média e covariância multivariada. Em casos em que *outliers* estejam presentes esta medida é distorcida. Dois métodos

podem ser usados para se obter uma estimativa multivariada robusta de médias e covariância. Um deles é o MVE (*Minimum Value Ellipsoid*), com mínimo efeito de *outliers* baseado em encontrar um hiperelipsóide que contenha um subconjunto de “boas” observações que minimizem o volume do elipsoide. Outro método é a estimativa *Minimum Covariance Determinant Estimation* (MCD), minimiza o determinante da matriz de covariância baseado em um hiperelipsóide gaussiano simétrico.

- A ACP é discutida, mapas de escores podem ser úteis para o entendimento dos processos geoquímicos.

- A escala de medida pode ter efeito significativo na ACP, a covariância entre elementos reflete a magnitude dos elementos, em consequência elementos com valores grandes tendem a dominar a matriz variância-covariância, isto tende a aumentar a significância destes elementos na ACP. A matriz de correlação representa as correlações interelementos, que são os equivalentes padronizados da matriz variância-covariância. Quando a matriz de correlação é utilizada todos os elementos passam a ter igual representatividade e as combinações lineares dos elementos são baseadas somente em suas correlações e não em seus valores absolutos. Estimativa robusta tanto de covariâncias como de correlações na ACP fornece uma melhor estimativa das médias das variáveis pela diminuição dos pesos da influência das observações anômalas.

- Em casos em que ocorrem *outliers* ou observações atípicas ou distribuições são não normais pode-se adotar certos procedimentos, como:

    - Se a distribuição marginal for censurada, encontrar um valor para substituir de modo que média e variância sejam boas estimativas da população, o que pode ser feito substituindo-se por um valor igual a ½ ou 1/3 do valor censurado, ou usando procedimentos estatísticos para estimar um valor de substituição com base em características estatísticas da porção não censurada.

    - Se *outliers* estiverem presentes pode-se removê-los para o cálculo de médias e covariâncias, ou aplicar procedimentos robustos que minimizem ou eliminem os efeitos destes valores.

    - Se a distribuição for não normal se pode fazer uma mudança da distribuição usando procedimentos estatísticos ou separações categóricas, como tipo de rocha, por exemplo, ou aplicar uma transformação que traga a distribuição para uma outra próxima da normal.

- Os pesos do modo R são os *eigenvectors*, que são escalonados de acordo com a raiz quadrada dos *eigenvalues*. No caso prático apresentado os pesos do modo Q foram interpolados pelo uso da krigagem.

- O objetivo principal da *Cluster Analysis* é a identificação de grupos naturais no conjunto multidimensional de dados, os métodos podem ser divididos a grosso modo entre hierárquicos e não hierárquicos. Os métodos hierárquicos se baseiam na união de variáveis (modo R) ou de observações (modo Q) por meio de medidas de similaridade, estes métodos assumem uma ligação constante de variáveis, o que nem sempre é uma assertiva razoável. O coeficiente de correlação (modo R) é a medida de similaridade mais comum em *Cluster Analysis*, a distância euclidiana pode ser usada como uma medida de proximidade por meio da qual as observações podem ser agrupadas.

- No caso do modo Q o tamanho da matriz de similaridade que contém a medida da distância métrica entre pontos pode se tornar tão grande que o tratamento se torne quase intratável. Métodos de origem arbitrária são não hierárquicos e podem oferecer alguma vantagem sobre os hierárquicos uma vez que os agrupamentos são formados com base nas similaridades multivariadas (proximidades) em vez de por coeficientes de correlação individuais. Estes métodos iniciam com uma quantidade inicial de grupos com um centro que podem ser especificados ou aleatoriamente escolhidos, e cada observação é alocada em um dos grupos com base na sua proximidade com os centros deles, em processo interativo, os centros dos grupos mudam até que seja encontrada uma solução estável.

- *K-Means Clustering* inicia com alguns centros de grupos, a distância entre cada observação e o centro é medida e a amostra é provisoriamente associada àquele grupo, este processo interativo continua até que se encontrem centros estáveis. O método requer uma escolha inicial do número de grupos. Se o número de centros for muito grande haverá grupos pequenos com poucas observações, se o número for muito pequeno a estrutura dos dados pode não ser realizada. Uma desvantagem do procedimento é que um agrupamento menor do que o ótimo pode resultar se os centros inicialmente determinados não ficarem dentro de grupos distintos. É comum aplicar agrupamento não hierárquico aos pesos da ACP.

- O uso da matriz de covariância é uma ferramenta para distinguir *background* de anomalias, a matriz de covariância contém informação sobre a variabilidade dos elementos e também sobre as suas interrelações.

- *Outliers* podem ser distinguidos da população de *background* pela distância de Mahalanobis de cada observação em relação ao centro do grupo. As distâncias podem ser comparadas com distâncias “esperadas” de uma distribuição multivariada normal, com uso do teste de qui quadrado.

- Dados geoquímicos podem ser interpretados utilizando-se uma sequência de passos, como segue:

    - Examinar cada elemento com histogramas, *box plots*, Q-Q *plots*, *scatter plots* e tabelas

    - Mapas de elementos mostrando intervalos de valores com símbolos.

    - Censurar a distribuição dos *outliers* mais elevados, ou extremos

    - Investigar os *outliers* para cada elemento (erro analítico, ou valor atípico?)

    - Ajustar os valores censurados se for o caso.

    - Se medidas de associação forem exigidas (correlação, covariância) durante a estimativa robusta, então transformar cada elemento usando *Box-Cox* e adotando os percentis 95-98, o que depende de uma exame visual prévio dos Q-Q *plots* ou dos histogramas, e aplicar transformações logarítmicas aos dados.

    - Examinar os *scatter plots* e os Q-Q *plots* para verificar a presença de múltiplas populações.

- Análise Exploratória de Dados Multivariados, esquema proposto:

    - Criar uma matriz para os dados transformados, atentando para *trends*/associações.

    - Usar estimativas robustas para computar médias e covariâncias para detectar *outliers*.

    - Aplicar técnicas de redução de dimensionalidade de dados, como ACP.

    - Gerar mapas dos pesos da ACP.

- Usar *Cluster Analysis* para isolar grupos de observações com características similares e observações atípicas.

- O uso de gráficos de Distâncias de Mahalanobis (D^2) aplicados a dados transformados pode ajudar a isolar *outliers* com base na seleção de elementos de interesse, mapas de distâncias, como >percentil 95, por exemplo, pode ser útil.

Os autores citam que o uso de fractais pode salientar padrões multivariados e *trends*, bem como a integração de resultados de estatística multivariada com geoestatística.

Galuszka, A. – 2007 - Different approaches in using and understanding the term "geochemical background" – practical implications for environmental studies. Polish J. of Environ. Stud. Vol. 16, No. 3, 389-395

O termo *background* geoquímico foi originalmente introduzido por exploradores em geoquímica na metade do século XX para diferenciar entre a abundância de um elemento em formações de rochas não mineralizadas e mineralizadas. Nas últimas décadas esta expressão se tornou mais crucial em ciências ambientais, ela é algumas vezes utilizada para distinguir valor antropogênico (poluição) de valor natural (geogênico).

Os métodos indiretos (estatísticos) englobam várias técnicas, como análise de regressão, gráficos de probabilidade, método fractal, mas, acima de tudo, técnicas usadas para eliminar *outliers* que são considerados antropogenicamente influenciados.

Métodos estatísticos de avaliação do *background* são criticados pelos geoquímicos por negligenciarem os processos naturais que influenciam a distribuição dos elementos ou compostos químicos nos materiais ambientais e por não considerarem a incerteza dos diferentes estágios, como amostragem, preparação da amostra e análise química, ou seja, a preocupação se dá mais com o número do que com a grande variedade de fatores que influenciam as concentrações nas amostras ambientais.

O método proposto estabelece que sejam tomadas como referência amostras coletadas em áreas nativas, como parques nacionais, áreas preservadas, restringindo a amplitude de dados obtidos na análise estatística, selecionando ecossistemas que garantam não ter sofrido influência antropogênica. Este método está sendo introduzido na Polônia e em Luxemburgo e parece ser bem promissor

para estudos de geoquímica ambiental. As vantagens mais importantes deste método são:

- é simples e claro de entender e executar
- leva em consideração o conhecimento do técnico sobre o comportamento dos elementos no ambiente, gerando maior robustez e melhor avaliação do *background*
- usa a abordagem estatística, influenciando uma maior precisão da avaliação do *background*
- não incorpora nenhum impacto antropogênico sobre a área estudada
- permite avaliar as faixas de *background* nas escalas regional e local, incluindo a especificidade da região

O autor conclui que existe uma forte necessidade de se fazer um arranjo terminológico de forma a estabelecer uma definição precisa para a expressão *background* geoquímico em ciências ambientais em contraposição à geoquímica exploratória. O autor sugere que a expressão ***background* geoquímico** seja usada apenas em geoquímica exploratória enquanto que **concentrações de *background*** seriam usadas em ciências ambientais.

Backgrounds, anomalism and geochemical targets – Internet (Geochemical Anomalies.mht) – 2008.

- Um nível de *background* de um elemento é normalmente considerado como sendo um intervalo típico de valores encontrados em certa litologia não mineralizada em uma certa área, na realidade o background seria aquilo que se gostaria que determinada rocha fosse.
- Uma anomalia é criada somente quando uma componente de interesse é adicionada a um *background* arbitrário a partir de alguma fonte externa, e não há necessariamente uma relação real entre um valor de *background* e um valor anômalo.
- Trabalhos regionais de geoquímica de sedimentos de corrente em fontes de diferentes litologias em geral recebem um mesmo *background* para fins estatísticos de tratamento, só é aceitável isto se considerarmos que a mistura de material durante o transporte “normalise” a variabilidade.

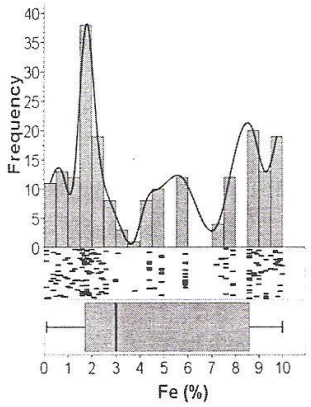
Carranza, E. J. M. - 2009 - Geochemical anomaly and mineral prospectivity mapping in GIS / Handbook of Exploration and Environmental Geochemistry, vol. 11, Elsevier, M. Hale editor - capítulo 3: exploratory analysis of geochemical anomalies, p. 51 – 84.

- A aplicação da estatística clássica fundamentalmente assume que os dados se constituem de amostras independentes e que têm distribuição normal; na realidade os conjuntos de dados geoquímicos unielementares invariavelmente contêm mais de uma população, cada uma delas representando um processo, pode-se afirmar que dados geoquímicos invariavelmente não seguem a distribuição normal e que não são espacialmente independentes. Neste caso, o uso de (média ± dois desvios-padrão) podem levar a resultados espúrios na determinação de anomalias geoquímicas.

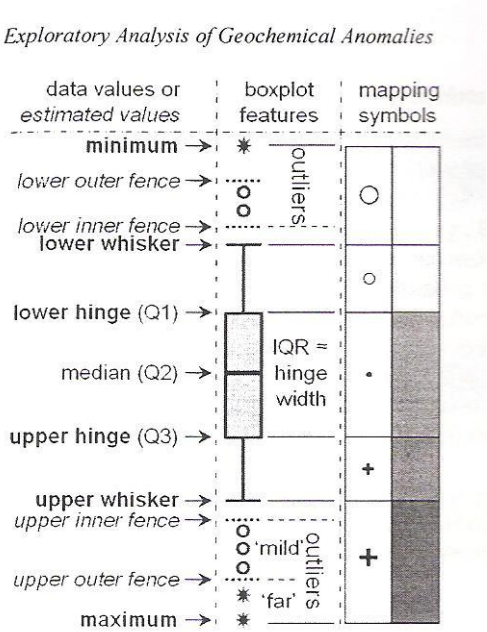
- No final dos anos 1970, Tukey (1977) introduziu o paradigma de EDA - *Exploratory Data Analysis* - para analisar dados que não seguem um modelo normal.

- 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 (b) descobrir a estrutura dos dados (c) definir variáveis significativas nos dados (d) determinar *outliers* e anomalias (e) sugerir e testar hipóteses (f) desenvolver modelos prudentes (g) identificar o melhor tratamento e interpretação possível dos dados. A estatística clássica e probabilística são abordagens confirmatórias, EDA é uma abordagem exploratória. EDA usa estatística descritiva e gráficos robustos que são quantitativamente distintos dos da estatística clássica, os produtos gerados na EDA são baseados nos dados mesmos e não em modelos de distribuição, como a normal, por exemplo.

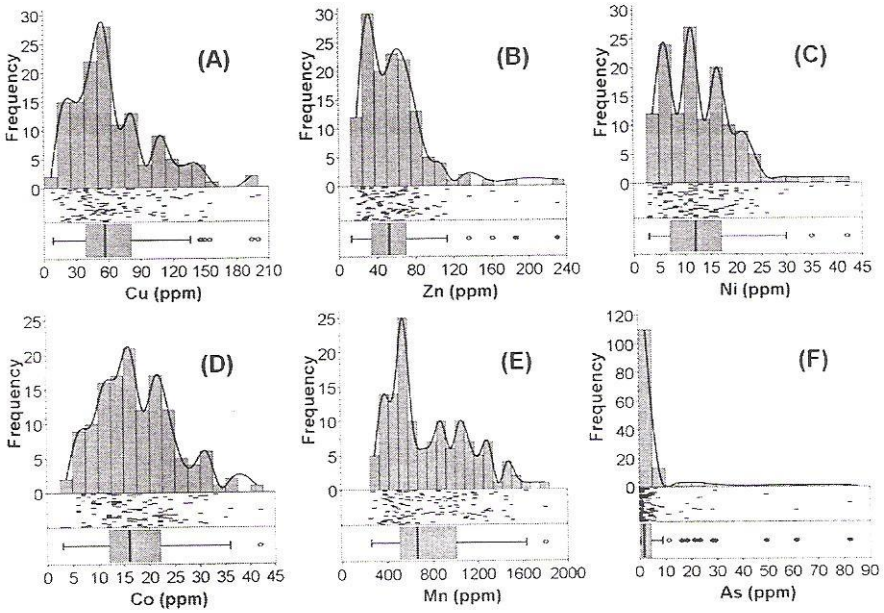
- Há três tipos de gráficos em EDA, chamados *density trace*, *jittered one dimensional scatterplot* e o *boxplot*, que são utilizados juntamente com o histograma. *Density trace* é similar ao histograma, mas ele descreve a densidade de distribuição empírica de dados univariados de uma maneira muito mais realística e sua forma não muda significativamente com a mudança do número de classes. O número apropriado de classes do histograma pode ser controlado pelo *jittered one dimensional scatterplot*, no qual os dados univariados são plotados em posições aleatórias através de uma banda estreita, usualmente com variação entre zero e um, ortogonal ao eixo de dados univariado. *Boxplot* é provavelmente o gráfico mais aplicado na EDA.



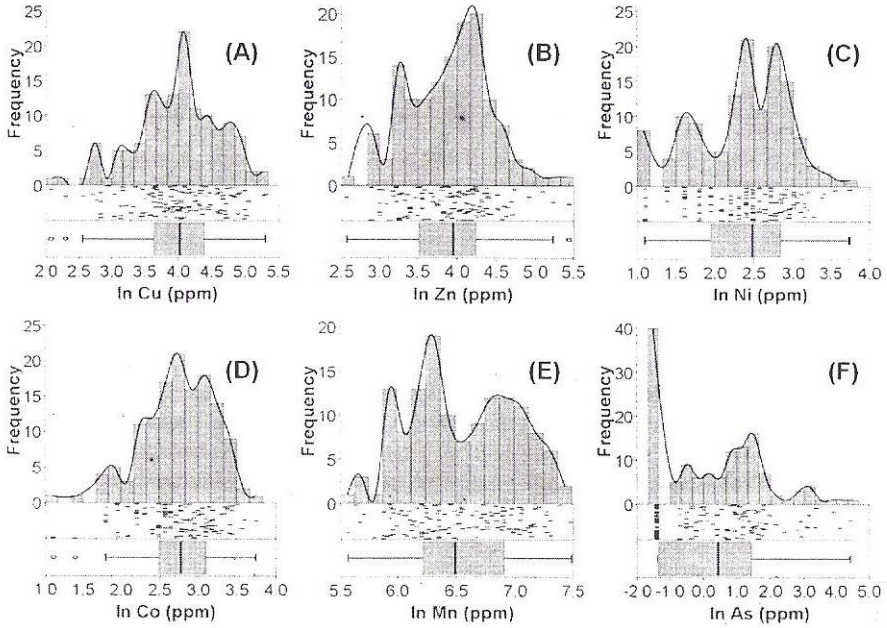
Histograma e gráficos EDA (*density trace, jittered one dimensional scatterplot e boxplot*)



Feições do *boxplot* representando características de um conjunto univariado. Textos em itálico representam valores estimados baseados no IQR ou largura da *hinge*, e textos em negrito representam valores nos quais um conjunto univariado pode ser dividido em cinco classes robustas.



Histogramas e gráficos EDA de distribuições de densidade empíricas de dados brutos unielementares, distrito Aroroy (Filipinas)



Histogramas e gráficos EDA de distribuições de densidade empíricas de dados logtransformados unielementares, distrito Aroroy (Filipinas)

- 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 = $\left| \textit{lower hinge - upper hinge} \right|$

- 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)

$$X_{LIF} = X_{LH} - 1{,}5 \times IQR$$

$$X_{LOF} = X_{LH} - 3 \times IQR$$

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:

$$X_{UIF} = X_{LH} + 1{,}5 \times IQR$$

$$X_{UOF} = X_{LH} + 3 \times IQR$$

Para dados logtransformados se usa os mesmos conceitos para os valores dos logaritmos dos dados.

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:

$$X_{LW} = \min (X[X > X_{LIF}])$$

e

$$X_{UW} = \max (X[X < X_{UIF}])$$

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 a *outliers*.

- Com base em um *boxplot* uma exploração unielementar de dados geoquímicos nos permite dividir os valores em cinco classes robustas, que são (a) mínimo - LW (b) LW - LH (c) LH - UH (d) UH - UW (e) 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 x 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.

- 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 que segue mostra comparações entre as diferentes determinações de *threshold* utilizando (média + 2 desvios-padrão), (mediana + 2 MAD) e o *boxplot* pelo valor UW, tanto para dados brutos como para logtransformados. Para dados brutos (mediana + 2 MAD) sempre é menor, seguido por *boxplot* (UW) e por (média + 2 desvios padrão), dependendo do conjunto de dados, e para dados logtransformados o *threshold* com (mediana + 2 MAD) é sempre mais baixo, seguido preferencialmente por (média + 2 desvios-padrão), sendo os do *boxplot* (com UW) geralmente os mais altos, dependendo do conjunto de dados.

	Mean+2*SDEV*		Median+2*MAD*		Boxplot *UW*	
	Raw	$Antilog_e$	Raw	$Antilog_e$	Raw	$Antilog_e$
Cu	139.72	184.93	96	120.30	136	200
Zn	121.76	139.77	88	108.85	113	187
Ni	26.31	34.81	22	26.58	30	42
Co	32.06	43.73	26	28.50	36	42
Mn	1461.93	1719.86	1120	1380.22	1630	1800
As	25.86	27.66	4	14.73	9.0	82.0
As*	30.89	29.96	7	11.94	11.0	48.9

Grunsky, E.C. – 2010 - The interpretation of geochemical survey data. Geochemistry: Exploration, Environment, Analysis, vol. 10, p. 27-74.

Problemas que comumente ocorrem em dados geoquímicos incluem casos de elementos que possuem distribuições “censuradas”, significando que há valores inferiores ao limite de detecção que somente podem ser reportados como menor do que, casos em que a distribuição dos dados é não normal, casos em que valores faltantes são reportados como zero, o que é diferente dos casos em que o valor é realmente zero em uma amostra, e combinações de dados com diferentes limites de detecção, além do problema da soma constante para dados composicionais. É perigoso assumir que que o fechamento dos dados (soma 100%) não tenha efeito sobre as medidas estatísticas geradas.

As *whiskers* são as linhas que se estendem além da *box*. Os extremos são marcados por barras verticais ao final das *whiskers*; alternativamente as *whiskers* podem se estender às *fences*. Que são definidas como o último valor antes de 1,5 x midrange, os valores que se situam além de 3 x *midrange* são plotados com símbolos especiais.

Q-Q *plots* são uma forma gráfica de comparar a distribuição de frequências, que é usualmente a distribuição normal, eles são equivalentes aos gráficos de probabilidades normais extensivamente usados por Sinclair (1976). A vantagem dos Q-Q *plots* é que cada observação individual é plotada e assim as características detalhadas dos grupos de observações podem ser observadas, eles também são úteis na identificação de valores extremos nas caudas das distribuições.

Alguns autores propõem grupos de valores em uma distribuição, como por exemplo a *upper fence* (Q3 + 1,5 x (Q3 – Q1)), a *lower fence* (Q1 – 1,5 x (Q3 – Q1)), os valores *lower outside* e *upper outside*, em que o valor 1,5 é substituído por 3, lembrando que Q1 e Q3 são o primeiro e terceiro quartis da distribuição dos dados, e (Q3 – Q1) é o *interquartile range*.

Um modo quantitativo de acessar a variabilidade é o uso de geoestatística (semivariogramas ou correlogramas), mas um problema é que os dados geoquímicos raramente apresentam estacionariedade, que significa invariância

locacional, ou seja, a relação entre os pontos é a mesma independentemente da localização geográfica. Assim, técnicas de interpolação como krigagem precisam ser utilizadas com precaução, particularmente se os dados cobrem diversos domínios geoquímicos nos quais o elemento apresenta características espaciais significativamente diferentes, assim o erro nas estimativas por krigagem tende a aumentar.

Estudos já mostraram que os métodos fractais podem ser usados para determinar *thresholds* de distribuições geoquímicas com base na relação espacial da abundância, onde a concentração de certo elemento por unidade de área satisfaz um modelo fractal ou multifractal a área do componente segue uma função *power* na sua relação com a concentração.

Outlier é definido como uma observação que é distintivamente diferente da distribuição da qual provém, se um *threshold* for definido então um *outlier* o excede, pode definir uma mineralização, como também pode representar um erro analítico. Uma anomalia é associada com um processo de interesse (alteração ou mineralização), enquanto que um *outlier* é um valor sem uma interpretação que requer mais investigação.

Quando um grupo de valores apresenta resultados maiores ou menores que os limites de detecção o efeito é chamado em inglês de *censoring*, muitas técnicas foram desenvolvidas para diminuir seus efeitos. De todo modo, os parâmetros da distribuição devem ser estimados com a parte não censurada dos dados, a substituição dos valores é chamada na literatura estatística de língua inglesa de *imputation*.

Se houver a presença de *outliers* ou mistura de populações a presunção de normalidade é violada. A aplicação de transformação de dados deve ser aplicada com cuidado para evitar o mascaramento da presença de múltiplas populações ou a presença de *outliers*.

Algumas transformações que podem ser aplicadas são a linear, que não modifica a forma da distribuição, embora o grau de dispersão (variância) possa ser afetada, também as transformações logarítmica, exponencial, *power* generalizadas *Box-Cox*, ou logaritmo decimal, elas modificam tanto a forma como as características da dispersão das distribuições e são as mais comumente utilizadas. Uma distribuição deve ser examinada para *outliers* antes e depois de transformações que venham a ser aplicadas.

Outro problema é o estudo de dados oriundos de diversas etapas de estudos, a uniformização deve partir do princípio de que um dos grupos deve ser o padrão, os outros sendo referidos a ele, são métodos difíceis de implementar e que exigem boa capacidade computacional.
Em situações em que há a presença de *outliers* ou observações atípicas há algumas opções a adotar:

- Se a distribuição marginal é censurada deve-se encontrar uma substituição de valores de modo que média e variância sejam bons estimadores dos seus correspondentes na população, o que pode ser feito por adotar certo valor, como ½ a 1/3 do valor mínimo, ou usar procedimentos estatísticos com substituição de valores com base em em características da porção não censurada.
- Se *outliers* estiverem presentes se pode removê-los todos para o cálculo de média e variância, ou aplicar procedimentos robustos que minimizem ou eliminem os efeitos destes valores.

Métodos de Cluster são divididos de modo genérico em hierárquicos e não hierárquicos, os hierárquicos são baseados na reunião de variáveis (Modo R) ou de observações (Modo Q) por meio de medidas de similaridade. O agrupamento hierárquico assume que que todas as variáveis são unidas em algum nível, o que pode não ser uma presunção razoável em alguns casos.

Métodos com origem arbitrária são não hierárquicos e podem oferecer algumas vantagens sobre os hierárquicos uma vez que os grupos são formados com base nas similaridades multivariadas em ves de por seus coeficientes de correlação individuais, estes métodos começam com um número inicial de grupos que pode ser especificado ou escolhido aleatoriamente, cada observação é alocada em um dos grupos com base em sua proximidade ao centro do grupo, em processo interativo que continua até que uma solução considerada estável seja encontrada.
É comum aplicar métodos não hierárquicos aos escores da PCA.

Outliers podem ser distinguidos da população de *background* pela determinação da distância de Mahalanobis de cada observação ao centro do grupo, se houver *outliers* então esta distância será maior que a esperada para os quantis por qui quadrado.

O autor apresenta uma estratégia para a análise de dados geoquímicos. Para análise preliminar coloca que se deve conhecer os dados, examinar cada elemento com histogramas, *box plots*, Q-Q *plots*, tabelas sumarizando os dados, investigação de *outliers* para cada elemento, ajuste de dados (*censoring*) se exigido, e para dados multivariados sugere criar matriz de dados brutos e transformados, aplicar técnicas de redução de dados (PCA, por exemplo) para identificar padrões e *trends* nos dados, usar mapas para os escores, aplicar métodos como Cluster para isolar grupos de dados com características similares.

Howart, R.J. & Garrett, R.G. – 2010 - Statistical analysis and data display at the Geochemical Prospecting Research Centre Group, Imperial College, London. Geochemistry: Exploration, Environment, Analysis, vol. 10, p. 289-315.

Estudos exaustivos sobre a aplicação das componentes principais robustas e a Análise Fatorial (AF) feitas por Turner (*Statistical Analysis of Geochemical Data Illustrated by Reference to thr Dalradian of NE Scotland, PhD Thesis, University of London*,1986) concluíram que a AF é preferível devido a que o uso de pequeno número de fatores força o agrupamento de variáveis, reduzindo a dimensionalidade do problema e aumentando a interpretabilidade.

Carranza, E. J. M. – 2011 - Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. J. of Geochem. Explor., 46 p.

Analisou dados de sedimentos de corrente de distrito epitermal a ouro em Aroroy, Filipinas, 135 amostras, -80 *mesh* (<177 µm), área total de 101 quilômetros quadrados, utilizou seis elementos (As, Cu, Zn, Ni, Co, Mn), análise por AA após decomposição com água régia. Usou As por ser *pathfinder* de Au, Cu/Zn/Ni/Co por serem bons indicadores de variações litológicas, e Mn por ser indicador de efeitos de óxidos e hidróxidos no ambiente superficial. Relata que 30% dos dados (40, exatamente) são censurados, ou seja, estão abaixo do limite inferior de detecção.

A transformação logratio (**alr**, **clr** ou **ilr**), comparada com a logtransformação, não melhora o mapeamento das anomalias dos elementos *pathfinder*, que refletem mineralização. Entretanto ao incluir os valores censurados com sua substituição por metade do valor inferior de detecção as transformações **clr** ou **ilr** podem apoiar a identificação de associações que refletem a mineralização. Variações nas concentrações de elementos traço em sedimentos de corrente são principalmente devidas a litologias e outros fatores não relacionados com mineralizações.

Um conjunto de dados de sedimentos de corrente é um exemplo de sistema fechado pois que contém variáveis composicionais que fazem parte de um todo. Três transformações foram feitas no estudo para “abrir” os dados, a *additive logratio* (**alr**), a *centered logratio* (**clr**) e a *isometric logratio* (**ilr**), das três somente a **ilr** produz representações apropriadas para sistema de números fechados no espaço euclidiano, no qual análise estatística pode ser aplicada.

As três opções utilizadas foram excluir do conjunto, o que ainda permite estudo multivariado, que exige **n > 10v**, onde **n** é o número de amostras e **v** o número de variáveis do estudo, em segundo substituir os valores censurados por 0,25 (metade do limite inferior de detecção para As) e em terceiro os valores de As foram substituídos pelas recomendações de Horn e colaboradores, por ser, à época, a proposta mais recente (Horn, K.; Templ, M. e Filzmoser, P. – imputation of missing values for compositional data using classical and robust methods, na Computational Statistics and Data Analysis 54, p. 3095 – 3107, que se baseiam na substituição por valores obtidos a partis dos **k** vizinhos mais próximos). Os dados foram então transformados para **ln** (logaritmo neperiano), **alr**, **clr** e **ilr**.

No estudo o autor adotou a C-A (*concentration area*) *fractal analysis* para separar áreas de *background* e anomalias tanto em elementos individuais como em multielementar. Para mapas de anomalias representando associações multielementares os conjuntos de dados transformados individualmente foram estudados por ACP. Os escores das PCs interpretados como refletindo mineralização foram submetidos à C-A *fractal analysis* para classificar e mapear anomalias representativas de associações multielementares. Os resultados desta aplicação foram sumarizados para mapas binários (*background* – anomalia).

Os resultados foram separados em três classes (*background*, baixa anomalia e alta anomalia), depois unificados reunindo as anomalias em um só grupo). Os mapas de *ln* e *alr* foram similares entre eles, e os de *clr* e *ilr* também entre eles. Estas três classes se mantiveram quando os dados censurados foram substituídos, caso em os quatro casos apresentaram resultados similares. Neste caso, em que os dados censurados foram substituídos por 0,25, o autor concluiu que a transformação **alr** apresentou resultados mais próximos da realidade conhecida da área quando da aplicação da análise fractal C-A.

Os dados transformados por **ln** e **alr** mostram, em aplicação da ACP, valores e escores das PCs tanto para conjuntos sem dados censurados como para dados censurados substituídos muito semelhantes, e os resultantes de **clr** e **ilr** da mesma forma quando os dados censurados recebem o valor de metade do limite inferior de detecção.

O autor conclui que como os dados originais não se conformam à distribuição normal, a aplicação de métodos unielementares de separação *background* – anomalia baseados em medidas de tendência central e dispersão tendem a ser errôneos. Conclui também que esta definição por análise C-A fractal supera os métodos de média mais dois desvios padrão e mediana mais dois MAD.

Conclui também que, para o caso em estudo, a transformação *ln*, excluindo ou substituindo os valores censurados, são inadequados para revelar associações multielementares que poderiam ser de interesse na exploração mineral, e também que a forte similaridade da CP 1 com dados transformados com *alr* e *ln* mostra que a transformação *alr* é inadequada para a abertura de dados composicionais, como é o caso dos dados de sedimentos de corrente.

Relata que as transformações *clr* e *ilr* são eficientes para a abertura dos dados composicionais e são adequadas para revelar associações multielementares que possam ser de interesse da exploração mineral.

Os resultados deste trabalho no que diz respeito a substituições de dados censurados não está de acordo com conclusões anteriores de estudos de Grunsky e Reinmann. Tendo em vista que as condições variam de área para área, a substituição de 30% dos dados censurados requer investigação em outras áreas em que for aplicada.

Lima, A.; Plant, J. A.; De Vivo, B.; Tarvainen, T.; Albanese, S. & Cicchella, D. - 2013 - Interpolation methods for geochemical maps: a comparative study using arsenic data from European stream waters. Geochemistry: Exploration, Environment, Analysis, vol. 8, p. 41 – 48

O autor demonstrou em estudos anteriores o valor de algumas novas técnicas de execução de mapas geoquímicos, incluindo interpolação multifractal e filtro fractal em escala regional, relata que estes mapas se mostraram superiores a outros gerados por métodos de interpolação como krigagem, por exemplo, distinguindo melhor anomalias e valores de fundo.

Foram gerados mapas com interpolação multifractal por MIDW (*multifractal inverse distance weighted*), que revelou distinção mais confiável entre *background* e anomalias, e por MWM (*moving weighted median*).

O estudo incluiu 26 países europeus, cobriu cerca de 4,5 milhões de quilômetros quadrados de área, malha baseada no *Global Terrestrial Network* (GTN), que cobre toda a Terra em células de 160 km de lado, foi dosado As em água.

Comparado com outros mapas gerados para a Europa ficou claro que técnicas convencionais como MWM e krigagem suavizam a variabilidade local dos dados, enquanto que MIDW cria um mapa geoquímico no qual a informação da estrutura local é retida, destacando as anomalias. Além disso, a krigagem gera grandes padrões anômalos ao redor de anomalias isoladas que mostram erroneamente grandes áreas da Europa como tendo elevados valores de As na água; assim, parece que a krigagem é o método menos adequado para gerar mapas geoquímicos regionais que possam ser interpretados com qualquer nível de confiabilidade. Mapas geoquímicos gerados por MIDW para objetivos ambientais mostram uma distribuição mais realística, que reflete tanto as variações litológicas como outros processos naturais que controlam as formações.

Kääriäinen, K. – 2016 - Reanalysis of the existing regional geochemical data around the Sakatti Ni-Cu-PGE target, Sodankilä, Finland. Master's thesis, University of Helsinki, 53 p.

Os dados trabalhados consistem de 1867 amostras amostradas em linhas separadas entre si por 1 a 1,5 km, com amostras a cada cem metros nas linhas, analisados Fe, Mg, Ca, Na, K, Ti, V, Cr, Mn, Co, Ni, Cu, Zn e Pb. Amostras preparadas por secagem a 70ºC, após o que foram peneiradas a < 0,06 mm. Foram testados alguns métodos de tratamento de dados, entre eles o coeficiente de correlação de *rank* de Spearman, já que os testes de normalidade mostraram que os dados não se ajustavam a este modelo. O coeficiente de Pearson exige

normalidade dos dados, o de Spearman não requer relação linear entre variáveis e pode ser usado para dados assimétricos. Na essência o coeficiente de Spearman é uma variação do de Pearson, a diferença reside no fato de o de Pearson ser medido diretamente sobre os dados e o de Spearman exigir que os dados sejam antes ranqueados para o cálculo.

Também foram utilizadas técnicas de ACP e Cluster K *means*. Existem vários algoritmos para Cluster, o hierárquico é um dos mais utilizados mas não é útil para conjuntos maiores que algumas centenas de amostras. Como o hierárquico necessita da matriz de distância ou de similaridade para todos os pares de dados em um conjunto ele se torna pesado computacionalmente para grandes conjuntos de dados, e em muitos casos os resultados são ambíguos. Considerando-se que o conjunto estudado tem 1867 amostras foi utilizado o método K *means*, que requer um número pré-definido de grupos, que pode ser determinado via determinadas abordagens.

Razões entre elemento podem ser usadas como impressões digitais do minério.
Self-organizing map (SOM) é uma implementação das redes neurais, forma imagens de dimensão baixa a partir de distribuições de dados de alta distribuição pela utilização de agrupamentos de dados, pode manejar grandes quantidades de dados e encontrar dependências subjacentes entre diferentes variáveis. No presente estudo o uso de SOM possibilitou encontrar quais amostras estavam mais relacionadas com mineralizações.

Na avaliação foi considerado que a ACP foi o método mais efetivo para o estudo das amostras de *till*, e tende a destacar amostras que têm concentrações elementares mais elevadas.

K-*means cluster*, entretanto, é sensitivo à presença de *outliers*. Os centros dos grupos tendem a ficar ao redor dos *outliers*.

SOM foi o método mais sensitivo para as limitações dos dados.

Iwamori, H.; Yoshida, K.; Nakamura, H.; Kuwatani, T.; Hamada, M.; Haraguchi, S. e Ueki, K. – 2017 - Classification of geochemical data based on multivariate statistical analyses: Complementary roles of cluster, principal component, and independent component analyses. AGU Publications, 19 p.

Este trabalho apresenta um novo método estatístico que efetivamente captura as estruturas de vários tipos de dados multivariados. O método se baseia na combinação de Análise de Cluster k-*means* não hierárquica (KCA), PCA e ICA (*Independent Component Analysis*). Eles são de abordagem não supervisionada,

e não requerem informação prévia e desvendam estruturas escondidas com base na classificação / agrupamento feitas somente sobre os dados.

KCA é um método simples e comumente usado para a partição dos dados em certo número de grupos, PCA e ICA descrevem os dados por meio de um conjunto de vetores que maximizam suas variância e sua não-gaussianidade, respectivamente, nos quais os dados são descritos em uma solução espacial contínua com escores individuais.

As razões para selecionar estes três métodos são:

- KCA e PCA são provavelmente os mais fundamentais ferramentais para a análise multivariada
- ICA não é tão comum como PCA mas é uma ferramenta única para a identificação de estruturas independentes escondidas

Os três métodos são recentemente pensados serem estreitamente correlacionados e podem ser integrados para uma análise efetiva dos dados.
No presente trabalho é descrita a relação destes três métodos para elucidar a estrutura total baseada principalmente em dados sintéticos, e é mostrada a utilidade potencial para geoquímica.

As variáveis observadas em uma suíte de amostras geralmente têm diferentes unidades de medida, diferentes amplitudes. Primeiramente uma padronização dos dados (a usual, cada dado menos sua média dividida pelo desvio padrão); muitos outros métodos de padronização podem, entretanto, ser utilizados para reduzir o efeito dos *outliers*.

Se os dados tiverem soma constante (ppm, por exemplo), então pré-processamento dos dados brutos tomando a *logratio* (transformação aditiva ou centrada) é um método útil, em vez de ou em adição à padronização primária (apenas com média e desvio padrão). Os dados padronizados então são analisados por KCA e PCA.

O princípio da ICA é direto: se as fontes independentes exibem distribuição não gaussiana, como distribuição assimétrica, um sinal de que a aleatoriedade mistura as fontes independentes seria mostrar uma distribuição próxima da gaussiana. Uma combinação linear de variáveis observadas seria maximamente não gaussiana se ela igualasse uma das ICs independentes.

ICA tem algumas poucas mas importantes limitações. A primeira é a de que as ICs precisam ter distribuições não gaussianas,, se tiverem distribuição gaussiana a ICA fracassa na identificação das ICs porque têm zero não gaussianidade e não podem ser discriminadas de outras componentes não independentes.

Um segundo ponto importante é que o usuário de ICA especifica o número de ICs a serem detectados, até um máximo igual à dimensão dos dados. No presente

trabalho se executou a PCA e se usou apenas as PCs que seriam escolhidas por aquele método. Este processo é conhecido pelo nome de redução da dimensionalidade e determina o número de ICs, que em muitos casos estabiliza a ICA por meio da redução de ruído. Assim, o número de ICs é determinado por um compromisso entre a redução da dimensionalidade e a retenção da informação original.

A solução da ICA é geralmente não única, o que requer tempo e dedicação para interpretação.

KCA e PCA para dados padronizados são úteis para identificar estruturas de dados controladas pela amplitude de variabilidade.

Em resumo, a combinação destes três métodos pode fornecer uma visão completa da estrutura dos dados.

Grunsky, E. C. & de Caritat, P. de – 2017 - Advances in the use of geochemical data for mineral exploration. In: Proceedings of Exploration 2017: Sixth Decennial International Conference on Mineral Exploration, p. 441 – 456.

Grunsky, E. C. & de Caritat, P. de – 2019 - State-of-the-Art of analysis of geochemical data for mineral exploration. In: geochemistry: exploration, environment, analysis, The Geological Society of London for GSL and AAG, 34 p.

Os problemas que são mais tipicamente associados com dados geoquímicos são soma constante, valores faltantes, censura de dados, reunião de dados, escalonamento de diferentes conjuntos de dados e adequada malha de amostragem.

Escalas locais ou de alta densidade variam de mais de cem a uma amostra por quilômetro quadrado, escalas regionais variam de uma por quilômetro quadrado a uma por cada quinhentos quilômetros quadrados, e escalas continentais podem variar de uma a cada quinhentos quilômetros quadrados a até uma amostra a cada cinco mil quilômetros quadrados.

Amostragem de material transportado (*till* glacial, sedimentos lagunares, sedimentos de corrente, colúvios ou alúvios) pode refletir quantidades variáveis de material transportado e mistura de vários processos, o que pode ser desejável.

Distinção entre grãos grosseiros (tipicamente > 63 µm e < 2 mm) e grãos finos (< 63 µm) é usada comumente, o material grosseiro pode ser considerado, em muitos casos, como representando localmente partículas derivadas ou minerais que não sofreram intemperismo, cominuição ou dissolução química, e a assinatura

geoquímica de matéria mineral de grão fino pode representar minerais que sofreram intemperismo, cominuição ou dissolução/precipitação química. A fração fina é geralmente considerada como refletora de uma gama mais ampla de processos geoquímicos, embora seja dependente da fonte do material e da natureza dos processos subsequentes que ocorreram. Certos métodos analíticos podem exigir determinadas frações granulométricas específicas.

A escolha do método analítico é crítica na interpretação dos resultados, a escolha da digestão é geralmente a mais importante. Diversos tipos de digestão ácida, incluindo quatro ácidos (HF – HCl – HNO3 – HClO4), água régia e numerosas extrações fracas/parciais dissolverão preferencialmente mineral específico, fases amorfas e orgânicas, ou íons adsorvidos. A digestão quatro ácidos é uma digestão "quase completa" que dissolve tudo menos os resistatos, como monazita, zircão e outros. O uso de água régia é útil para dissolver sulfetos e óxidos, deixando alguns minerais silicatados não afetados. Ácidos fracos tendem a dissolver o revestimento de grão de minerais e/ou espécies adsorvidas que são associadas com processos de alteração/mineralização. Métodos antigos mas ainda correntes de instrumentação incluem Espectroscopia de Absorção Atômica (AAS), *fire-assay* é o método preferido de preparação em minerais de minério para a determinação de Au, Pt e Pd.

Quando os dados são expressos como proporções há duas limitações: os dados precisam somar uma constante (100%, por exemplo), e se um valor muda um ou mais outros também devem mudar. A soma constante ou "*closure*" resulta em dados não confiáveis, o uso de logaritmos de razões (*logratio*) é requerido quando se mede momentos como variância/covariância.

Utilizar o limite de detecção ou alguma substituição arbitrária (como 0,5 do limite inferior de detecção, por exemplo) para dados censurados, embora seja de uso comum, pode gerar bias em cálculos estatísticos, especialmente multivariados.

Considerando que as componentes principais (PCs) dominantes identificam processos ativos, e as PCs últimas podem refletir processos subamostrados ou ruídos, o uso das PCs dominantes pode ser efetivamente usado para classificação apenas algumas poucas variáveis. Uma regressão da variável de interesse (Zn, por exemplo) contra a PC dominante refletirá a associação daquela variável com o processo dominante, para o bem da regressão se usa os dados brutos. Valores elevados de resíduo derivados da regressão de uma variável contra uma PC podem refletir a existência de uma associação potencial com mineralização. Estes resíduos elevados, quando plotados em mapa, podem realçar áreas que venham a ser alvos para *follow up* subsequente.

Bonham-Carter, G. F. e Grunski, E. C. – 2018 – Two ideas for analysis of multivariate geochemical survey data: proximity regression and principal component residuals. In: Daya Sagar, B. S.; Cheng, Q. & Agterberg, F.

Editors – Handbook of mathematical geosciences – fifty years of IAMG, Springer Open, cap. 23, p. 447 – 465

Proximity Regression **(PR)** é um método exploratório de predição de halos multielementares (ou vetores multielementares) ao redor de uma feição geológica, como um depósito mineral, por exemplo, usa a regressão múltipla diretamente para predizer a proximidade de uma feição geológica (a variável resposta) a partir de elementos geoquímicos (variáveis explanatórias).

Residual Principal Components **(RPC)** é outro método multivariado exploratório, após aplicar a ACP convencional um subconjunto de CPs é utilizado como variáveis explanatórias para predizer um elemento simples, separando o elemento em partes residuais e preditas para facilitar a interpretação.

A proximidade a feições selecionadas tem sido usada em análises multivariadas de várias formas, mas usualmente como função de ponderação e não como variável a ser predita diretamente.

Em trabalhos de geoquímica regional podemos estar interessados em entender quais variáveis são boas preditoras de proximidade a um depósito mineral ou a uma outra feição selecionada com locação conhecida. Isto é frequentemente referido em exploração mineral como o encontro de bons vetores da mineralização. Se a feição de interesse for um depósito mineral esta abordagem pode ser usada para encontrar novos depósitos.

É desejável remover o efeito de algum processo geológico (ou processos) que esteja refletido em uma ou mais CP, por exemplo, na análise de geoquímica de *till* a primeira CP é comumente interpretada como sendo devida ao efeito do transporte de *till*, assim pode ser desejável olhar a distribuição do elemento removendo a CP1. Entretanto há situações onde é útil examinar padrões espaciais de um determinado elemento após remover a CP1, ou várias CPs, isto pode ser alcançado pelo que se chama PCR. É uma regressão direta usano o elemento selecionado como uma variável resposta, e CP1 (ou combinação de CPs) como variáveis explanatórias.

Os resíduos (variável resposta observada menos variável resposta predita) fornecem a distribuição do elemento desejado após a remoção do efeito da CP1 (ou combinação de CPs). Se a CP1 é interpretada como sendo devida ao transporte então os resíduos representam os valores do elemento após a remoção do efeito do transporte do *till*.

RPC residual foi primeiramente utilizada pelo autor em um estudo de urânio em solos na bacia Athabasca.

Suponha que tenhamos uma matriz de dados geoquímicos com as linhas sendo amostras e as colunas sendo os elementos; adicionalmente temos as medidas de distância de cada amostra refletindo a menor distância de uma amostra a uma feição geológica (depósito mineral, intrusão, falhamento, etc).

Sendo X a matriz que recebeu a transformação **clr** para os elementos, com linhas como amostras e as colunas como os elementos. Os elementos geoquímicos são as variáveis explanatórias, e a coluna vetor Y contém os valores de proximidade, a variável resposta. A geoquímica é usada para "explicar" a resposta. Outros autores mostraram que as técnicas de dados multivariados distinguem rochas vulcânicas alteradas de não alteradas.

O método RPC é uma extensão direta da aplicação da PC, seguida por uma série de regressões lineares múltiplas. PC regular é feita na forma usual sobre a matriz de correlação calculada a partir das variáveis elementares que sofreram transformação **clr**. Aqui o objetivo é focar em um elemento selecionado para separar (ou partição) este elemento composicionalmente e espacialmente usando os resultados da PC.

Pode ser decidido predizer o elemento a partir da PC1, ou PC1 e PC2, ou PC1, PC2 e PC3, e assim por diante. Para cada uma destas seleções uma regressão múltipla é feita com as PCs selecionadas como variáveis explanatórias, e o elemento escolhido como variável resposta.

O caso prático apresentado consiste de 1611 amostras com 48 elementos, tanto maiores como traços. Antes da transformação clr todas as variáveis foram convertidas para ppm, PC foi feita sobre os 48 elementos. O objetivo é compreender melhor como o urânio é particionado entre dois granitos, o Nueltin e o Hudson. Regressões múltiplas foram feita (usando U-**clr**, e não U não transformado), começando comPC1, e depois adicionando PCs até 12. Para cada regressão U predito e U residual foram calculados e mapeados e um registro de ajuste foi definido. O resultado mostrou que PC1 não respondeu por muita variação de U, mas PC2 e PC3 mostraram significativos aumentos no grau de ajuste.

Destaca-se que o mapa de predição mostra um padrão fortemente correlacionado com o granite Neutlin, enquanto que o mapa de resíduos é fortemente correlacionado com o Hudson. As PCs 1 a 5 "explicam" o U no granito Neutlin, enquanto que a PC residual mostra onde U ocorre no Hudson. A RPC particionou urânio em duas partes que têm interpretações geológicas distintas.

O resultado não seria a probabilidade de ocorrência de uma ocorrência mineral, por exemplo, mas sim a proximidade predita à ocorrência mineral mais próxima.

Deve ser ressaltado que ao fazer a PCA sobre dados geoquímicos a transformação *logratio* é essencial, pois que o efeito de fechamento para introduzir artefatos nos resultados da PCA é bem conhecido. Experiência também tem

mostrado que a análise de resíduos exige que os dados geoquímicos elementares usados como variável resposta precisam receber a transformação **clr**, uma vez que os resultados da regressão são pobres se as variáveis resposta utilizadas não foram transformadas.

A análise de proximidade permite o uso de dados geoquímicos multielementares para predição direta da proximidade de feições geológicas diversas, a RPC é uma maneira útil de partição particular de elementos geoquímicos que pode facilitar a interpretação geológica.

Carr, J. R. – 2018 – Mathematical geology by example: teaching and learning perspectives. In: Daya Sagar, B. S.; Cheng, Q. & Agterberg, F. Editors – Handbook of mathematical geosciences – fifty years of IAMG, Springer Open, cap. 41, p. 831 – 847

Durante a década de 1930 a Análise de Componentes Principais (ACP) foi utlizada pelos psicólogos para agrupar pacientes de comportamentos similares, resultando em melhor entendimento deles. Três décadas depois os sedimentologistas (como Imbrie e Purdy em 1962 e Klovan em 1966) usaram ACP para agrupar amostras de sedimentos com base nas suas características sedimentológicas, permitindo indicações sobre fonte dos sedimentos, ambiente deposicional, composição ou outra condição importante para a interpretação geológica.

Pode a análise multivariada de dados ser ensinada sem uma prévia explicação, ou ao menos uma revisão, sobre a eigendecomposição dos dados? Se a resposta for sim somos levados a uma forma de ensinar que trata os algoritmos multivariados como uma caixa preta.

A significância relativa de cada fator com respeito à informação total é determinada somando-se os *eigenvalues* e então dividindo cada *eigenvalue* por esta soma para se obter uma proporção, os *eigenvectors* são os fatores.

Se aspectos de maior amplitude (regionais) forem de mais interesse que os aspectos de escalas menores então a suavização deveria ser usada durante a estimação para minimizar as escalas menores. Por outro lado, se desejarmos visualizar variabilidade espacial ao menor tamanho possível não se deveria minimizar a suavização durante a estimação.

De um modo geral regiões do mapa que mudam apreciavelmente quando os parâmetros de estimação são modificados são as regiões mais esparsamente amostradas.

A krigagem não é necessariamente o objetivo último da análise geoestatística. O variograma como ferramenta isolada tem uma variedade de utilizações que são independentes da estimação. Exemplos são muitos e incluem isolamento de

ruídos, classificação de texturas de imagens digitais e análise e modelamento fractal auto-similar.

Reimann, C.; Fabian, K.; Birke, M.; Filzmoser, P.; Demetriades, A.; Negrel, P.; Oorts, K.; Matschullat, J. & **de Caritat, P. – 2018 - The GEMAS Project Team - GEMAS: Establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. Applied Geochemistry 88, p. 302** – **318**

O projeto GEMAS (*Geochemical Mapping of Agricultural Soil*) coletou 2.108 amostras de solos no horizonte A, em 33 países europeus, cobrindo cerca de 5,6 milhões de quilômetros quadrados. A fração < 2 mm destas amostras foi analisada para 53 elementos por ICP-MS e ICP-AES, com digestão por água régia modificada (HNO_3 / HCl / H_2O), e os resultados foram utilizados para estabelecer as faixas de variação dos *backgrounds* e definição dos *thresholds*, com uso de técnicas estatísticas.

Concentrações elevadas em solos podem ser devidas à ocorrência de mineralizações, ou de tipos não usuais de rochas, como serpentinitos, folhelhos negros ou intrusões alcalinas, ou ainda podem ter sido causados por atividades humanas. Dependendo da biodisponibilidade estas concentrações podem apresentar riscos devido à toxicidade do elemento.

Várias formas de identificar os *thresholds* foram tentadas. Reinmann e colaboradores (2005) sugeriram ser possível utilizar Mediana ± 2 MAD, mais robusto que pelo uso de média acrescida/diminuída de dois desvios padrão. A desvantagem deste método é que se aplicado a dados brutos ele fornece valores de *threshold* muito conservadores (baixos), em geral ao redor do percentil 90, ou seja, produz uma grande quantidade de locais a serem checados. A razão é as distribuições de dados geoquímicos são muitas vezes fortemente assimétricos à direita; a abordagem correta ao se usar esta fórmula seria calcular esta fórmula sobre dados logtransformados em base 10 e então fazer a retrotransformação dos resultados e adotar *threshold* como sendo

Threshold = 10b

onde **b = (mediana dos logs + 2 MAD dos logs)**

esta abordagem fornecerá resultados bem comparáveis aos obtidos com o uso das *boxplots*.

Outra abordagem é o estudo dos gráficos de probabilidade acumulativos, que é uma ferramenta poderosa para detectar processos que causam desvios na distribuição geral. *Outliers* podem muitas vezes serem detectados nas quebras destes gráficos.

A *International Organization for Standardisation* (ISO, ano 2005) propôs que para detector *outliers* em um trabalho de geoquímica de solos se adote as proposições de Tukey (*boxplots*). Este método tem muitas vantagens, é baseado na *boxplot* (introduzida originalmente por Tukey em 1977) e depende apenas da distribuição dos dados e não de modelos pré-estabelecidos, permite a definição de um *threshold* para *outliers* mesmo que nenhum deles esteja presente no conjunto de dados. Este método pode ser considerado como um dos mais confiáveis e poderosos para o cálculo de valores significativos de *threshold* para qualquer conjunto de dados.

Também o percentil 98 foi usado como critério definidor.

Concluem os autores que os três métodos adotados que responderam bem aos objetivos foram as curvas acumulativas nos gráficos de probabilidade, o percentil 98 e as *boxplots*, independente da escala do trabalho.

Precisa ser registrado, entretanto, que valores de *threshold* completamente diferentes foram obtidos para diferentes áreas, o que reflete a larga diversidade de condições geológicas em escala europeia. Estes valores de *threshold*, com base na distribuição das concentrações de *background*, não serão capazes de separar contaminação antropogênica de concentrações naturalmente elevadas, devidas à ocorrência de depósitos minerais ou variações litológicas. Além disso, não fornecerão qualquer indicação de riscos potenciais de toxicidade à saúde humana ou ao ambiente. Nestes casos uma avaliação local específica será necessária, preferencialmente com *thresholds* derivados de dados ecotoxicológicos.

Há casos de influências relacionadas a condições climáticas locais e a zonas de vegetação e não relacionados com a geologia. Poucas anomalias relacionadas com fontes antropogênicas podem ainda ser detectadas nesta escala e nesta densidade de amostragem. Exemplo são as grandes cidades (casos de Ag e Hg em Londres, Ag, Hg, Pb em Paris, Hg em Kiev), agricultura (Cu em vinhedos no sul da Europa, em especial no NW da Alemanha), e algumas, mais pontuais, anomalias relacionadas com a vizinhança de plantas de beneficiamento ou indústrias.

As recomendações, em termos de *follow-up* na escala europeia, com base nos dados do projeto, seriam adensar a malha de amostragem ao redor das grandes cidades e/ou outras fontes de emissão conhecidas (plantas de beneficiamento, indústrias, fundições), também unir e nivelar os trabalhos existentes em muitos locais de alta densidade de amostragem de trabalhos de solos existentes, junto aos serviços geológicos dos países membros. Também continuar o estudo com

outras fontes de amostragem (rocha, água, plantas) e análises multielementares delas.

Yazdani, M.; Alinia, F. Geochemical Anomaly Separation by Statistical and Fractal Methods in the Sehezar Valley of Tonekabon – 2019 - Northern Iran. *Preprints*, 2019100237

Para identificar a área anômala foram utilizadas 71 amostras de sedimentos de corrente analisadas por ICP-MS. Os dados foram normalizados e sobre eles aplicados alguns métodos, como análise univariada, método PN, análise bivariada e análise multivariada. O corpo intrusivo mais importante é a massa granítica Sehezar.

Na abordagem univariada primeiramente os dados foram normalizados e se aplicou o teste não paramétrico de Kolmogorov-Smirnov, que ajudou a determinar a normalidade da população. Histogramas e gráficos Q-Q foram feitos. Utilizou média mais dois desvios padrão para determinar o background.

O método PN de determinação de anomalias é baseado em dois princípios, o primeiro é o aumento no valor da variável e outro é o aumento na sua frequência relativa. Assim, a intensidade de cada anomalia é dependente dos dois fatores, com a possibilidade de aparecer uma amostra ao valor desejado (P), quanto menor a possibilidade mais intensa a anomalia, e com o número de amostras tomadas (N), quanto menos forem mais intensas serão as anomalias. A multiplicação dos dois fatores, chamada NxP, pode ser usada como um critério para seleção de anomalia. Obviamente, quanto menor este valor for em relação à unidade mais intensas são as anomalias, pois no modo normalidade a multiplicação do número de amostras com um teor assumido em uma certa probabilidade de ocorrência deste teor será unitária. O valor P para cada elemento em cada amostra é igual à probabilidade de ocorrência do maior ou igual valor da variável investigada na amostra. As prováveis anomalias têm o valor PxN entre 1 e 0,1, e as anomalias mais intensas tiveram valor menor que 0,1.

O método fractal concentração – área foi utilizado, é um método de separação de anomalia e background, que é um dos métodos mais fortes com um fator de segurança. É baseado no comportamento fractal das distribuições geoquímicas na natureza, na quantidade de área que certo teor em particular ocupa na área estudada. Quanto mais o teor de um elemento aumenta menor tende a ser a área ocupada.

Se o teor de cada contorno for igual a **v**, podemos fornecer uma equação para materiais para focar nas propriedades fractais:

$$A(\geq V) \propto V - \alpha$$

A(≥V) é a área acumulada incluída pelas linhas de contorno do mapa, em que o correspondente teor é maior ou igual a **v,** a quantidade **α** na realidade representa a dimensão relacionada aos diferentes domínios **V**. Plotando as áreas contra os teores numa gráfico logarítmico podemos calcular a dimensão de cada comunidade por meio do declive das linhas. Pontos de quebra (mudança na inclinação da linha) neste gráfico representam mudanças de uma comunidade para outra, servem para separar background de anomalias com um grau variado de um elemento e até em certos casos maiores e menores mineralizações relacionadas a um e outro. Mudanças de um grupo a outro representam as condições geológicas, geoquímicas e mineralógicas.

Os autores concluem que os métodos estatísticos usados neste projeto mostraram resultados similares.

Chen, D.; Wei, J.; Wang, W.; Shi, W.; Li, H. & Zhan, X. – 2019 – Comparison of methods for determining the thresholds of geochemical anomalies and the prospecting direction – a case of gold deposits in the Gouli exploration area, Qinghai province. Minerals 2019, vol. 9, n. 368, 23 p.

As anomalias geoquímicas têm um papel importante na exploração mineral, os autores analisaram 764 amostras de sedimentos de corrente em 125 quilômetros quadrados, escala 1:50.000, e usaram como exemplo. Usaram as técnicas de frequência cumulativa, análise de singularidade local e EDA para determinar os *thresholds* respectivos de ouro e determinar o *threshold* abnormal. ACP robusta foi usada para explorar as associações elementares. Singularidade local foi a melhor técnica no caso para delinear as anomalias geoquímicas fracas, que não foram delineadas pelos outros dois métodos.

Na conferência Internacional de Sydnei em 1976 o *threshold* abnormal foi definido como um valor determinado pela geoquímica de acordo com alguns testes e análises, ele pode identificar as anomalias associadas com mineralizações. Uma definição que passou a ser utilizada é a que define como sendo um valor que distingue dados abnormais de dados não abnormais. Com o decréscimo da intensidade das anomalias isto se torna mais importante.

Há muitos métodos para determinar o *threshold* e as anomalias:

- Método estatístico tradicional
- Método gráfico direto
- Método da seção geoquímica
- Método da distância mertensítica
- Cálculo do elemento individualmente

- Método da frequência cumulativa
- Análise da singularidade local
- EDA
- Outros

Os primeiros cinco objetivam principalmente dados geoquímicos de distribuição normal, e os métodos de frequência cumulativa, análise de singularidade local e EDA são aplicáveis aos dados em geral, sem considerar a influência dos *outliers*.

Para ouro os valores variaram de 0,5 a 8,4 ppb, média de 1,32 ppb, mediana de 1,1 ppb, desvio padrão 0,95, variância 0,91, assimetria 3,46 e curtose 17,57. Sabe-se que assimetria em torno de zero e curtose em torno de três aproximam a distribuição de uma curva normal, mas a distribuição de ouro na área não se conforma a esta distribuição. Mesmo assim os autores acharam razoável usar o *threshold* uniforme abnormal aos dados.

Método da frequência cumulativa – classifica um grupo de dados do menor para o maior, e toma a parte dos dados na parte do topo como abnormal para estudo. É conveniente usar a teoria de probabilidades e a estatística para dados com distribuição normal, mas não para dados com distribuição geral, e o ouro, como já referido, não apresenta distribuição normal nesta área.

Análise de singularidade local – a essência da geometria fractal está na auto-similaridade e na invariância de escala. Auto-similaridade se refere à similaridade entre partes e o todo na forma, na função, na informação, na estrutura e em outros aspectos, e a invariância de escala se refere a poder-se alargar ou reduzir o tamanho apropriadamente sem alteração da estrutura do objeto. A análise de singularidade local é um importante progresso para modelamento fractal e multi-fractal de dados geoquímicos. A singularidade pode ser estimada a partir das concentrações elementares em pequenas vizinhanças com base na seguinte equação:

$$\mathbf{X = c \times r^{\alpha} - E}$$

onde **X** representa a concentração do elemento, **c** é um valor constante, **α** é a singularidade, **r** é uma medida de distância normalizada, como o lado de uma célula, e **E** é a dimensão euclidiana. O método exige muitos cálculos para encontrar os valores.

EDA – *exploration data analysis*, usa *boxplots* para interpretar a distribuição dos dados e determinar o *threshold* abnormal, este método não requer que os dados (brutos, após processamento de cosseno positivo, cosseno inverso, ou logaritmo) obedeçam à distribuição normal.

ACP robusta – método muito efetivo para identificar associações, entretanto, por ser dado tipicamente composicional, tem o efeito de fechamento dos dados, o que

pode causar correlação virtual dos dados originais no processo de ACP. Para diminuir este efeito os dados geoquímicos podem ser “abertos” no pré-tratamento. Atualmente os métodos mais utilizados para tal são o *log-ratio* aditivo (alr), o *log-ratio* centrado (clr) e o *log-ratio* isométrico (ilr). Carranza, E. J. M. (2011, *Analysis and mapping of geochemical anomalies using logratio transformed stream sediment data with censored values*, J. Geochem. Expl., vol. 110, p. 167 – 185)) comparou estes métodos. Por ser a seleção do denominador artificial na **arl**, a soma de todas as variáveis após a transformação **clr** é zero e os dados resultantes são colineares, de modo que a transformação **alr** e a **clr** não são indicadas para a ACP e a **ilr** é. A transformação **ilr** resolve o problema da colinearidade dos dados resultante da **clr**, preservando suas vantagens, como isometria entre a parte e o todo. As dimensões dos dados no **irl** se reduzem a apenas uma, o que dificulta a interpretação, mas combinando com pesos e escores se consegue interpretar os resultados de modo a refletir a presença de mineralizações.

Os *thresholds* obtidos foram 1,9 para frequência cumulativa, 1,7 para singularidade local e 2,55 para EDA. O *threshold* abnormal pelo método de frequência cumulativa foi 1,9 ppb, e tem certa artificialidade e empirismo. Por EDA se obteve a maior área acima do *threshold*, áreas anômalas não são contínuas e são espalhadas.

Em termos comparativos os resultados de frequência cumulativa e EDA têm melhores resultados em áreas de anomalias óbvias, mas para áreas de anomalias fracas os resultados não são tão significativos como os do método de singularidade local. Este método não apenas destaca as anomalias de ouro como também delimita bem as anomalias fracas, os autores concluem que em áreas de *backgrounds* geológicos complexos ele é uma ferramenta poderosa para identificar anomalias fracas, assim consideradas aquelas de pequeno contraste.

Grunsky, E. C. & de Caritat, P. – July 2019 - State-of-the-art analysis of geochemical data for mineral exploration. In: Geochemistry, Environment, Analysis, Geological Society of London, número 210161

Os autores fazem uma revisão do estágio do conhecimento até o ano de 2019 na interpretação de dados de geoquímica em exploração mineral.

A densidade de amostragem é um aspecto crítico do planejamento de uma campanha geoquímica e subsequente interpretação. Densidade local ou de alta densidade tem amostras entre um e cem por quilômetro quadrado, trabalhos de escalas regionais entre uma e quinhentas amostras por quilômetro quadrado, e serviços de escalas continentais têm malhas de uma amostra a cada 500 até 5.000 quilômetros quadrados.

A escolha do material a ser amostrado também tem grande importância, citam que a amostragem de *bedrock* revela processos geoquímicos *in situ*, amostragem de regolito que é derivado por intemperismo *in situ* do *bedrock* pode apresentar uma assinatura geoquímica que reflete tanto o protólito quanto seu intemperismo. A amostragem de material transportado (*till* glacial, sedimentos lagunares, sedimentos de corrente, sedimentos de *overbank*, materiais de colúvio ou alúvio pode refletir vários tipos de transporte e mistura de diversos processos, o que pode ser desejável em algumas situações.

Em meios amostrais que apresentam mistura de de matéria mineral com matéria orgânica a fração analisada pode ser importante para distinguir entre os processos geológicos e geoquímicos. Uma distinção entre fração grossa (tipicamente >63 µm e <2 mm) e fração fina (<63 µm) é usada comumente. A fração mais grossa pode ser considerada, em muitos casos, como representante de partículas derivadas localmente, ou minerais que não sofreram intemperismo, cominuição ou dissolução química, e os de granulometria fina podem representar minerais que sofreram estes efeitos, eles são considerados de um modo geral como representativos de uma gama mais ampla de processos geoquímicos. Também deve ser levado em conta que a granulometria pode ser exigência de certos métodos analíticos, que necessitam cominuição a certas faixas granulométricas.

A escolha do método analítico a ser adotado, que inclui a digestão da amostra e subsequente instrumentação analítica também desempenha papel muito importante. A escolha da digestão é geralmente a mais importante, há diversos tipos utilizados comumente, como quatro ácidos (HF – HCl – HNO_3 – $HClO_4$), água régia (HCl – HNO_3), além de diversas extrações fracas (ou parciais), que preferencialmente dissolvem um mineral específico, ou uma fase orgânica ou amorfa, ou focam em certas características físicas (íons absorvidos, cátions intercambiáveis).

A digestão a quatro ácidos é uma caracterizada por ser “quase completa”, que dissolve tudo exceto os minerais resistatos, como zircão, monazita e outros.

O uso de água régia é útil para dissolver sulfetos e óxidos, deixando a maior parte dos silicatos não afetada.

Extrações com ácidos fracos tendem a dissolver os revestimentos de grãos de minerais e/ou espécies absorvidas que são associadas com processos de alteração e mineralização.

Outros métodos de preparação de amostras incluem uma fusão total em vez de digestão na qual material fino é misturado de forma a um disco ou algo que possa então ser analisado diretamente ou tomado em solução com um ácido. A solução ácida resultante é então levada a um instrumento analítico após a diluição apropriada. Entre as técnicas estão ICP-OES e ICP-MS, nas quais a digestão ácida é primeiramente aspirada em uma câmara e convertida a um plasma em alta temperatura. Depois um espectro de emissão ótica é produzido, onde cada

elemento tem um único espectro de emissão (ICP-OES), ou um espectrômetro de massa pode ser usado para separar os elementos ou as moléculas com base na sua assinatura de massa específica (ICP-MS).

Métodos antigos de instrumentação, mas ainda correntes, incluem Espectroscopia de Absorção Atômica (AAS). *Fire assay* é o método preferido de preparação em materiais de materiais de minério na determinação de Au, Pt e Pd. Fluorescência de raios X (XRF) e ativação neutrônica (INAA) têm a habilidade de analisar uma amostra sem digestão úmida, permitindo assim uma análise total.

Quando os dados são expressos como proporções, surgem duas limitações: a primeira é a de que os dados são restringidos a um determinado espaço de variação e podem ter soma constante (100%, por exemplo), e em segundo lugar quando um valor muda outros também têm que mudar para manter a soma total constante. Este problema não pode ser solucionado apenas selecionando partes. O problema "soma constante" ou "*closure*" resulta em medidas estatísticas não confiáveis, e as relações entre os elementos em geoquímica são controladas por leis naturais.

A presença de dados censurados também merece atenção, usando o limite ou algum valor arbitrário relacionado a ele pode gerar enviesamento nos cálculos estatísticos.

No que diz respeito ao tratamento dos resultados há técnicas multivariadas não supervisionadas, como ACP, ICA (*Independent Component Analysis*), escalonamento multidimensional (MDS), e processo baseados em modelos , como Cluster. Um problema essencial a ser contornado é o problema do "*closure*". A transformação **clr** (*centred logratio*) é uma transformação útil para dados geoquímicos. ACP com dados **clr** transformados é feita com dados ortonormais, ou seja, estatisticamente independentes, e pode refletir processos lineares associados com restrições estequiométricas.

Os autores aplicaram técnicas de interpretação em duas áreas de estudo, Nunavut, Península Melville (Canadá) e Thomson, Austrália.

Discutem algumas proposições ao se estudar dados de programas geoquímicos:

- As CPs dominantes refletem combinações lineares de elementos e suas relações, controladas por estequiometria mineral. Relatam que os resultados refletem bem o conhecimento geológico disponível nos casos de aplicação.

- Processos subamostrados podem representar processos alteração e mineralização associadas com um tipo específico de depósito.

- Uma regressão de um mineral de interesse confrontado com uma CP dominante pode refletir a associação daquele elemento com o processo dominante. A escolha feita, de se usar dados brutos, se baseia na criação de segurança para a regressão, os minerais são variáveis independentes. Não há nenhuma facilidade de executar uma regressão de um elemento com qualquer dos espaços de transformação (**alr, clr, ilr**). Cada transformação a presenta diferentes problemas em qualquer análise envolvendo regressão.

- Altos valores residuais derivados da regressão de um elemento confrontado com a CP dominante pode refletir processos que são potencialmente associados com mineralizações. Estes resíduos elevados, quando plotados em um mapa, podem destacar áreas para follow-up de exploração.

Os autores concluem que o uso de métodos e técnicas multivariadas de alta qualidade, associado com técnicas analíticas avançadas, e em escalas regionais e de detalhe, são a nova etapa futura na busca de províncias minerais do tipo *greenfields*.

Yin, B.; Zuo, R.; Xiong, Y.;Yongsheng, L. & Yang, W. – 2021 – Knowledge discovery of geochemical patterns from a data-driven perspective. Journal of Geochemical Exploration 231, 9 páginas

Os autores referem a utilização dos métodos de Indicadores Locais de Associação Espacial - *LISA (Local Indicators of Spatial Association)*, Análise de Componentes Principais e *Deep Autoencoder Network (DAN)* na exploração de associações espaciais de padrões geoquímicos, na extração de associações entre elementos e na detecção de anomalias geoquímicas relacionadas à mineralização de Au – Sb no distrito Daqiao, província chinesa Gansu (o distrito está localizado no domínio leste do Orógeno West Qinling, uma das maiores regiões prospectivas para ouro na China, o depósito Daqiao tem mais de 105 toneladas de Au com teores entre 3 e 4 g/t). Referem que as anomalias geoquímicas identificadas por meio da técnica DAN exibiram forte correlação espacial com locações de depósitos minerais conhecidos e pode fornecer uma orientação significativa para mais exploração mineral no distrito.

Foram estudadas 2.090 amostras de sedimentos de corrente coletadas como parte de um levantamento geológico regional na escala 1:50.000 acompanhado de estudos de recursos minerais, foram analisados 15 elementos, sendo sete (Cu, Pb, Zn, Cd, W, Mo, e Au) analisados por ICP, dois (Ag e Sn) por espectrometria de emissão, dois (Ba e Mn) espectrometria de emissão plasmatômica e os demais (As, Sb, Bi e Hg) por espectrometria de fluorescência atômica.

Os autores apresentam um breve histórico de algumas abordagens, como a estatística de Moran, introduzida para acessar o padrão de distribuição espacial de objetos ao se estudar as relações entre processos contínuos e discretos distribuídos em duas ou mais dimensões, também discorrem sobre a ESDA (*Exploratory Spatial Data Analysis*), uma extensão da *Exploratory Data Analysis* (EDA, de Tukey, 1977), que quantifica a associação espacial em um conjunto de dados pela utilização de coeficientes de autocorrelação espacial do tipo Moran e Geary, abordagens que focam na estatística global e não podem acessar as instabilidades locais de cada observação do conjunto de dados.

Moran I é um método efetivo de identificação de *clusters* locais e de *outliers* locais presentes nos dados, valor positivo indicando que a feição apresenta vizinhos com valores similares do atributo, sejam eles elevados ou baixos, e valores negativos representando dissimilaridade com relação à vizinhança, representando um *outlier*. Em outras palavras, é feita uma distinção entre as presenças de *clusters* e *outliers*. Outra situação abordada é o fechamento dos dados, que pode resultar em correlações espúrias entre variáveis geoquímicas, citam as proposições de transformações denominadas *alr (additive logratio)* e *clr (centered logratio)* para lidar com o fechamento dos dados composicionais. A transformação *ilr (isometric logratio)* representaria corretamente os dados composicionais no espaço Euclidiano, mas mostram falta de correspondência entre as variáveis originais e as transformadas por *ilr*, o que dificulta a interpretação, geralmente se adota retro-transformação dos dados gerados.

O objetivo principal da DAN é garantir que a saída reconstruída seja muito próxima dos dados iniciais pela minimização dos erros de reconstrução. O reconhecimento das anomalias geoquímicas baseadas no uso do DAN repousa na construção de erro maior das amostras geoquímicas anômalas do que as amostras de background, pois que amostras de pequeno tamanho são geralmente ligadas a probabilidades de detecção mais baixas. Esta reconstrução de erro é o índice de anomalia do DAN, que se baseia no princípio de que amostras grandes têm um erro de reconstrução menor do que as pequenas ao se diferenciar anomalias geoquímicas de *background*. Após a seleção de parâmetros a reconstrução dos erros de cada célula, os escores geoquímicos de anomalias, é estimada pelo uso do DAN.

A taxa de sucesso foi estabelecida por comparação das células com depósitos minerais conhecidos, as medidas obtidas medem o quão áreas geoquimicamente anômalas se ajustam a depósitos minerais conhecidos. Estes resultados obtidos demonstram que DAN pode acomodar problemas não lineares importantes, permitindo a identificação de anomalias geoquímicas relacionadas com mineralizações e mesmo reconhecendo padrões geuímicos escondidos. Ainda mais importante é o fato de que DAN revela padrões de anomalias escondidas em áreas de diferentes tipos de depósitos minerais.

Os autores concluem que as anomalias geoquímicas obtidas pela aplicação de DAN são fortemente associadas com as localizações de mineralizações

conhecidas, podendo, então, fornecer informações críticas para a continuação da exploração mineral de certas áreas.

Zuo, R.; Wang, J.;Yihui, X. & Ziye, W. – 2021 – The processing methods of geochemical exploration data: past, present, and future. Journal of Geochemical Exploration 132, 9 páginas

Neste trabalho os autores apresentam uma revisão do estado da arte de métodos populares de processamento de dados gerados por exploração geoquímica e para a identificação de anomalias geoquímicas relacionadas com mineralizações.

Dados de exploração geoquímica, como dados composicionais, enfrentam o problema do fechamento dos dados *(closure problem)*. É preciso inicialmente "abrir" os dados por meio da *logratio transformation*.

A Estatística Clássica inicialmente utilizava a análise univariada pelo uso de média e desvio padrão, alguns pesquisadores utilizavam mediana e desvio mediano absoluto como alternativa para reduzir a influência da presença possível de *outliers* nos dados, determinando um *threshold* mais robusto. A transformação Box-Cox e a transformação logarítmica foram empregadas no pré-processamento de dados de exploração geoquímica de modo a fazer com que eles se ajustassem a uma distribuição normal, permitindo a aplicação dos conceitos da Estatística Clássica. Contrastes também foram utilizados para a determinação de *threshold*, *background* e para isolar anomalias geoquímicas. Os gráficos de probabilidade, que envolvem as distribuições teóricas e definem objetivamente o *threshold* foram largamente adotados para separar subpopulações misturadas e para identificar anomalias geoquímicas relacionadas com mineralizações.

A Estatística GAP determina o threshold e a probabilidade correspondente de que valores acima do *threshold* pertençam à população de *background*.

Stanley and Sinclair (1989) compararam os gráficos de probabilidade e a Estatística GAP na seleção de *thresholds* para dados de exploração geoquímica e concluíram que ambos os métodos exigiam pressupostos sobre a distribuição teórica dos dados. A Estatística GAP original envolve apenas a distribuição de frequência dos dados geoquímicos originais e ignora a variabilidade espacial das concentrações geoquímicas dos elementos. Posteriormente pesquisadores chineses desenvolveram uma Estatística GAP Extendida, um híbrido entre o método original e e as médias móveis, esta técnica identifica anomalias locais em configurações geoquímicas variadas e fornece confiabilidade estatística ao *threshold* determinado.

Exploratory data analysis (EDA) foi proposta por Tukey (1977); o *boxplot*, que é uma técnica popular da EDA é usualmente aplicada para explorar a estrutura interna dos dados de exploração geoquímica e para determinar o *threshold*.

Exploratory spatial data analysis (ESDA) é uma extensão da EDA, tem por objetivo visualizar e explorar dados geoespaciais pelo uso de diversos métodos, como nuvens de semivariograma/covariância, nuvens de covariâncias cruzadas, mapas de Voronoi, histogramas, *Q-Q plots* e *trend analysis*. Também se faz a utilização do índice de Moran I, que mede a associação espacial de cada observação com seus vizinhos, é uma abordagem efetiva na identificação de *clusters* locais e *outliers* espaciais. O gráfico de Moran é utilizado para visualizar e avaliar as feições espaciais de uma observação simples e mostrar tipos locais de autocorrelação espacial para cada observação.

Estatística multivariada define medidas matemáticas para entendimento das interdependências entre as variáveis e sua relevância no problema sob estudo, aí incluídas a ACP (Análise de Componentes Principais) e a Análise de Fatores (AF), duas das mais populares técnicas multivariadas utilizadas no estudo de dados de exploração geoquímica; reduzem a dimensionalidade do problema, mas são sensitivas à presença de *outliers*. Mais recentemente apareceram usos da ACP robusta (que reduz a influência dos outliers, que substitui a tradicional matriz de covariância pelo uso do determinante da mínima covariância) e da ACP ponderada espacialmente.

Geoestatística, relacionada a problemas de interpolação, fornece dados não enviesados e otimização de estimativas, e também tem sido utilizada em combinação com a teoria multifractal e com a inferência bayesiana.

O conceito fractal permite retratar um padrão geométrico que se repete em diferentes escalas, ligado à auto-similaridade, gera a dimensão fractal (D). Alguns autores revelaram a natureza fractal das paisagens geoquímicas. A *multifractal inverse distance weighting model (MIDW)* foi proposta para superar as deficiências do método do inverso das distâncias.

Machine Learning (ML) é um campo da Inteligência Artificial (IA), tem a capacidade de quantificar padrões complexos e não lineares, sem requerer modelos prévios de obediência dos dados a alguma distribuição de probabilidades. É dividida tradicionalmente em supervisionada e não supervisionada, ambas com aplicações conhecidas em exploração geoquímica.

Dados de exploração geoquímica estão sujeitos ao problema de fechamento (*closure problem*), a soma total em uma amostra tem soma constante (1 ou 100%). É conhecida a aplicação da *additive logratio (alr)* e da *centered logratio (clr)*, transformações que tratam deste problema. Também é conhecida a *isometric logratio (ilr)*, transformação que fornece uma correta representação dos dados composicionais no espaço euclidiano. Estes métodos, entretanto, também têm suas desvantagens, a seleção do denominador na *alr* é subjetiva, a *clr* gera resultados sujeitos à multicolinearidade por ser a soma das variáveis transformadas para *clr* igual a zero, a transformação *ilr* perde uma variável, e as variáveis geradas na transformação *ilr* são de difícil explicação.

APLICAÇÕES DE GRÁFICOS DE PROBABILIDADE EM EXPLORAÇÃO MINERAL APRESENTAÇÃO RESUMIDA

Alastair J. Sinclair, Probabbility Graphs in Mineral Exploration, The Association of Exploration Geochemists, 1976, 95 p.

- 700 amostras de solo são analisadas para Cu, variando seus valores entre 10 e 100 ppm. A variável (Cu em solo) é contínua entre estes limites porque, pelo menos teoricamente, qualquer valor intermediário poderia ser assumido por uma amostra. Na prática, é claro, um valor como 927,341 nunca será assumido, por exemplo. Em contraste às variáveis contínuas estão as referidas como discretas, ou seja, aquelas variáveis que tomam somente valores específicos. Por exemplo, o número de minerais em uma seção polida de um minério é uma variável discreta. Ela pode ser 1, 2, 3, etc minerais em uma dada seção mas não 1,372 tipos minerais.

- É importante notar a diferença entre os usos estatístico e geológico do termo amostra. Uma simples amostra de solo é uma amostra em um contexto geológico, mas é simplesmente um elemento na amostra estatística.

- Algumas vantagens dos histogramas como meio de representação visual de dados são:

    1) a variação total dos dados em uma amostra é aparente,
    2) a moda pode ser reconhecida facilmente,
    3) o intervalo de maior abundância de valores pode ser estimado rapidamente, e
    4) a forma geral da densidade de distribuição de dados é aparente.

- Uma vantagem adicional de um histograma é que o agrupamento preparatório de dados fornece uma forma relativamente conveniente para calcular a média e a variância pelo método dos dados agrupados.

- Há vários pontos concernentes à construção de um histograma que são dignos de menção. Primeiro está a escolha do intervalo de classe. Segundo Shaw (1964) um intervalo de classe é melhor escolhido entre ¼ e ½ do desvio padrão dos dados. Se o intervalo de classe é muito grande a forma verdadeira da distribuição dos dados é mascarada – se muito pequena então muitas lacunas aparecem no histograma resultante.

- Em geral, em construção de histogramas é boa prática incluir uma listagem de 1) título, 2) N, o tamanho da amostra, 3) o intervalo de classe, e 4) a média e o desvio padrão dos dados.

- A distribuição normal de Gauss foi primeiro anunciada como uma teoria de medida de erro. Por exemplo, nós podemos querer testar a reprodutibilidade de um método químico de análise de solos. Uma amostra simples pode ser dividida em 10 sub-amostras, cada uma delas analisada usando o mesmo método. Os 10 valores obtidos não serão necessariamente exatamente os mesmos devido a variações randômicas no procedimento analítico. A distribuição dos valores medidos ao redor da média segue o que é conhecido como a distribuição de densidade Gaussiana ou normal dada pela seguinte fórmula:

$$y = (1 / \sigma\sqrt{2\pi}) \cdot e^{-1/2(X - \mu)2 / \sigma2}$$

onde μ é a média aritmética, X é qualquer medida e σ^2 é a variância da população. A expressão gráfica é a curva em forma de sino.

- Na sua forma conceitual mais simples a distribuição lognormal é vista como uma distribuição normal dos logaritmos (em qualquer base) de um grupo de dados. Tem sido descrita em detalhe por Aitchison e Brown (1957) e pode formar uma base para a teoria de erros multiplicativos assim como a distribuição normal é a base de uma teoria de erros aditivos. Por exemplo, uma estreita aproximação à lognormalidade é mostrada comumente por variáveis tais como 1) elementos menores em geoquímica, 2) muitas variáveis geofísicas, 3) teores e tonelagens de depósitos minerais, 4) dados de tamanho de sedimento, 5) capacidade de reservatórios d'água, 6) tamanhos de reservatórios de óleo, e assim por diante.

- Talvez o maior problema em conciliar um modelo lognormal com muitos dados reais é encontrado com distribuições polimodais. Onde as populações componentes não se sobrepõem cada uma pode ser examinada individualmente para lognormalidade. A freqüência de ocorrência de tais distribuições lognormais polimodais indica que é lógico esperar que populações que se sobrepõem também se aproximem de modelos lognormais.

- Papel de probabilidade cumulativa padrão, comumente referido como um papel de probabilidade aritmética, tem uma escala aritmética (a escala da ordenada na maioria dos papéis que há nos EUA) e uma escala de abcissa de percentagem não usual (ou probabilidade). Ordenada e abcissa comumente são reversas no papel de probabilidade fora dos EUA e nos usados por sedimentologistas. A escala de probabilidade (% cumulativa) é arranjada como uma distribuição de densidade normal plotada como uma linha reta.

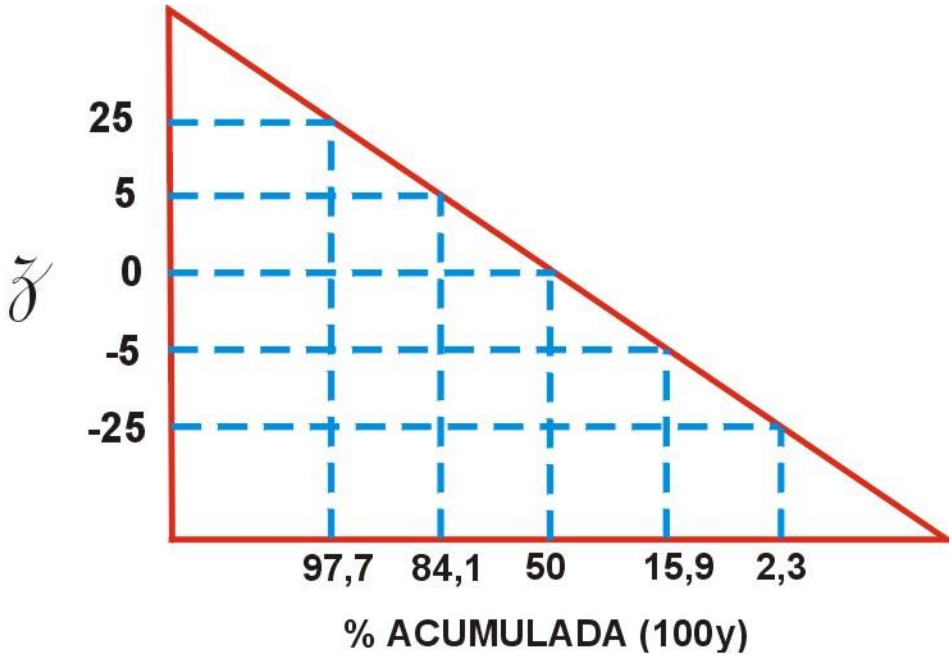


Diagrama ilustrando o relacionamento direto entre valores de Z de uma distribuição normal padronizada e % cumulativa da escala de “probabilidade”

- Está aparente que a ordenada é aritmética em termos de “números de desvios padrão” de um valor central de referência, zero, que corresponde ao valor médio da distribuição. Uma linha reta mergulhante (qualquer linha) é então introduzida como um meio de definir a escala de probabilidade por projeção dos % cumulativos conhecidos da escala de probabilidade. A % cumulativa a ser assinalada em qualquer ponto projetado sobre a escala de probabilidade é o valor 100y da equação

$$y = (1 / \sqrt{2\pi}) \text{ Integral(de } -\infty \text{ a } z) \; e^{-z2/2}$$

onde y representa a proporção de área (ou valores) em uma distribuição de densidade normal padrão que está abaixo do valor especificado padronizado “Z”.

- É aparente que valores de % cumulativa sobre a escala de probabilidade são diretamente proporcionais aos valores de z. Consequentemente, escalas logarítmicas podem ser determinadas simplesmente por multiplicação dos valores de z por qualquer constante apropriada.

- Um 2º tipo de papel de probabilidade usado comumente é o papel de probabilidade logarítmico. Papéis de probabilidade logarítmica e aritmética têm escalas de probabilidade comparáveis; o 1º, entretanto, tem uma escala logarítmica para o 2º eixo enquanto que o 2º tem uma escala aritmética. A escala logarítmica tem base 10.

- Na literatura os gráficos de probabilidade são apresentados usando a escala de probabilidade tanto como ordenada como abcissa, e as percentagens cumulativas começando tanto pelos valores mais baixos como pelos mais altos de um certo grupo de dados. Assim, há 4 maneiras diferentes de apresentação, todas elas têm sido usadas por vários autores. Neste trabalho a escala de probabilidade é escolhida como abcissa porque a maioria dos papeis-gráficos de probabilidade encontrados comercialmente nos EUA são construídos assim. Em adição, frequências (probabilidades) são acumuladas de altos para baixos valores. Conforme Lepeltier, isso evita que o último ponto seja não plotável (% cumulativa 100) no terminal de alto valor, que está comumente no intervalo de maior interesse em dados de exploração mineral. Além do mais, há instâncias em que tal procedimento fornece um ponto adicional para plotagem comparado com o procedimento reverso de acumulação dos baixos valores para os altos.

- A seção precedente mostra que uma população simples, cumulativa, normal, plotada sobre um intervalo de probabilidade cheio define uma linha reta sobre um papel aritmético por causa da natureza da escala de probabilidade. Similarmente, percentagens cumulativas de valores aritméticos de uma população lognormal definem uma linha reta sobre um papel log probabilístico. Inversamente, percentagens cumulativas de logaritmos de uma população lognormal definem uma reta quando plotamos sobre um papel de probabilidade aritmético.

- Na prática se plota precisamente os mesmos dados que seriam usados para construir um histograma cumulativo. Os pontos assim obtidos seriam somente coincidentemente plotados exatamente sobre uma reta. Em geral teria uma certa quantidade de espalhamento pelo erro amostral esperado
- O valor médio de cada população pode ser lido como a ordenada correspondente ao percentil 50. Quando as populações são normais (linha reta sobre papel de probabilidade) os valores da média, mais e menos um desvio padrão, podem ser estimados quase precisamente pelos valores das ordenadas que correspondem aos percentis 16 e 84 respectivamente ($M_{aritmética} \pm S$).

- Na prática, uma estimativa de 2S é obtida com a diferença positiva entre os valores dos percentis cumulativos 84 e 16. Esta diferença é dividida ao meio para estimativas. Note que média e desvio padrão de distribuições normais são citados neste manual como $M_{aritmética} \pm S$, uma forma que não seria confundida com a representação do erro padrão da média.

- Em geral, para dados plotados em mesma escala, um padrão linear fortemente mergulhante indica um desvio padrão grande se comparado a um padrão linear suavemente mergulhante.

- No trato com distribuições lognormais, duas aproximações são possíveis:

1) Plotar valores logarítmicos sobre papel de probabilidade aritmético

2) Plotar valores não transformados sobre papel log-probabilístico.

- 1º método é usado em Sedimentologia, onde frações granulométricas dos sedimentos são cotadas em valores phi (logaritmos em base 2 de diâmetro de peneira). Estimativas de parâmetro (X e S) é dada na maneira descrita antes. Estas estimativas, entretanto, se referem a log-valores.

- Comumente, com grande quantidade de valores é demorado e inconveniente transformar valores aritméticos em logarítmicos, em cujo caso os dados aritméticos cumulativos são plotados diretamente sobre um papel de log-probabilidade. Esse procedimento evita referência a tábuas de logaritmos porque a transformação é fornecida graficamente e automaticamente na plotagem, como um resultado da escala de ordenada logarítmica.

- Os parâmetros estimados lidos de um gráfico linear sobre um papel de log probabilidade são antilogs de:

1) média aritmética dos logaritmos dos valores

2) média aritmética dos logaritmos mais um desvio padrão, e

3) média aritmética dos logaritmos menos um desvio padrão

- Consequentemente, o valor médio determinado é a média geométrica dos dados originais e os 2 valores circundantes que cercam aproximadamente 68% de valores na distribuição são locados assimetricamente ao redor dessa média geométrica. Neste manual a forma adotada para designar distribuições lognormais será a média geométrica seguida entre parênteses por antilogs de:

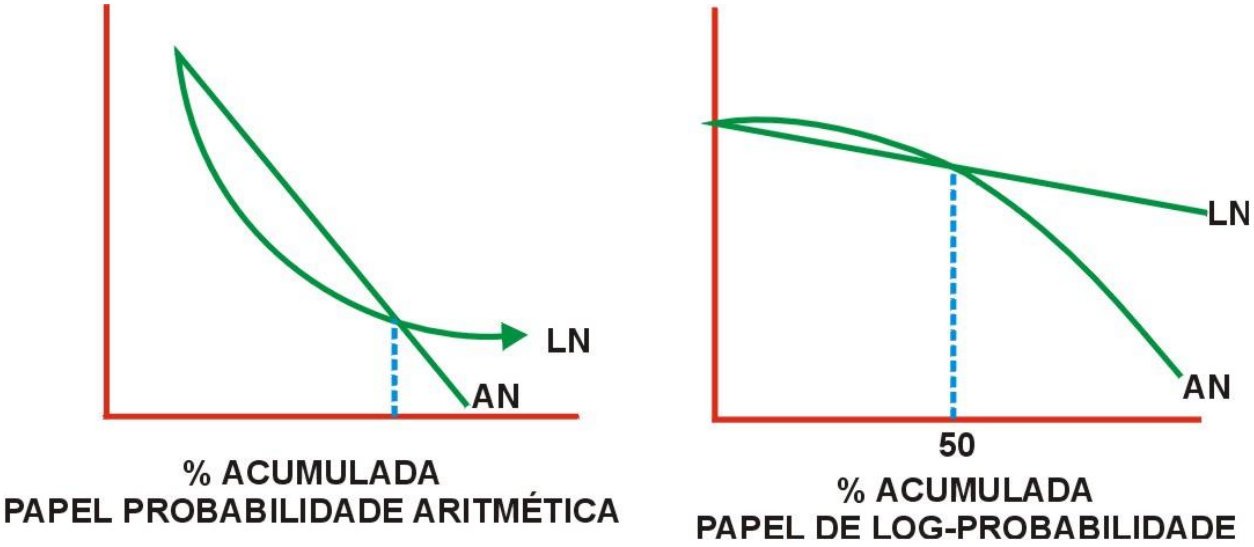
1) média de logaritmos mais um desvio padrão, e

2) média de logaritmos menos um desvio padrão

- Como um exemplo, consideramos a população 100 (250, 40). Logaritmos destes 3 valores são 2, 2,3979 e 1,6021. Estes são equivalentes a $2 \pm 0{,}3979$ e a simetria da distribuição em unidades log se torna aparente.

- É útil conhecer as formas das curvas cumulativas que resultam se uma população normal é plotada inadvertidamente sob papel de log-probabilidade (ou vice-versa).

- Considerando uma distribuição aritmética normal (AN) com parâmetros 100 ± 50. A mesma distribuição sobre papel de probabilidade log-normal dá um padrão curvo pronunciado, côncavo para baixo, aparente. Note que a maior curvatura está no sentido de valores baixos. Essa curva (AN) tem semelhança geral com uma distribuição lognormal truncada no topo, discutida mais adiante; mas, como regra, as duas podem ser distinguidas por se notar se a maior curvatura está no sentido dos valores altos ou baixos.



- Sumarizando brevemente: distribuições lognormais plotadas em papel de probabilidade aritmética produzem padrões curvos com concavidade para cima; distribuições normais plotadas em papel de log-probabilidade produzem padrões curvos com concavidade para baixo. Em geral gráficos com concavidade para cima são assimétricos em direção a valores altos, e vice-versa.

- Uma distribuição truncada é definida como uma distribuição de densidade, tanto normal como lognormal para nossos propósitos aqui, para a qual todos os valores acima ou abaixo de algum valor particular não são representados em dados úteis. Uma distribuição truncada simples pode ser descrita como truncada no topo ou truncada na base dependendo de qual terminal da distribuição está faltando. Distribuições truncadas não devem ser confundidas com distribuições censuradas. Uma distribuição censurada é uma para a qual medidas são conhecidas para itens sobre um lado de um valor específico mas somente o número de itens é conhecido no outro lado. Exemplos comuns de distribuições censuradas são grupos de dados analíticos para os quais alguns valores são registrados como

a) maior do que um valor particular, ou

b) zero ou não detectados ou menor do que o limite de detecção; se os dados censurados são ignorados no cálculo dos percentuais acumulados o gráfico resultante sobre o papel de probabilidade é idêntico ao da distribuição truncada.

- Distribuições truncadas surgem por uma série de razões, tanto naturais como artificiais. Consideremos, por exemplo, um sedimento com distribuição lognormal de tamanhos de partículas. Separação durante coleta de amostras pode remover todas as partículas abaixo de um tamanho específico deixando uma distribuição truncada na base dos grãos remanescentes.

- Em certos casos altos valores analíticos têm sido propositadamente ignorados no registro de dados levando a uma distribuição truncada no topo do ponto de vista do observador. O efeito de truncação de uma distribuição lognormal pode ser melhor visto pelo exame de distribuições hipotéticas truncadas. Aqui, uma população simples com parâmetros 50(118,22) foi recalculada assumindo que várias proporções têm sido removidas de sua parte superior (truncada no topo) e inferior (truncada na base). Em ambos os casos as curvas mudaram assumindo que 10, 25 e 50% da população simétrica original está faltando.

- Note as feições gerais dessas curvas:

1) Pronunciada curvatura ou afastamento de um padrão linear está mais evidente sobre o intervalo de % acumulada que foi truncado.

2) A curvatura produz um “aplainamento” da distribuição cumulativa no terminal que foi truncado.

- Em uma situação real representando uma distribuição truncada, o reconhecimento de truncação de topo ou de base é evidente do terminal da curva de acumulação na qual ocorre aplainamento. É possível, usando uma tentativa e procedimento de erro, estimar quase precisamente a percentagem da população que está faltando. Isto é dado pela replotagem da curva “dados reais” assumindo várias proporções que estão faltando. Sabendo que menos que 50% de uma população está faltando, uma grosseira estimativa da % que falta é obtida no ponto onde a curva começa a aplainar pronunciadamente.

- Logo se assume que uma população truncada pode ser reconhecida pela forma da curvatura do gráfico de probabilidade. Na prática, alguma ambigüidade pode existir.

- Em dados de exploração mineral algumas distribuições truncadas surgem artificialmente como resultado de serem alguns dados ignorados.

- Por outro lado seria possível que tais padrões surgissem naturalmente como resultado de uma amostra com *bias* de uma população .

- Poderíamos imaginar, por exemplo, uma grande zona geoquimicamente anômala na qual abundantes valores-picos não ocorressem na área amostrada. Na seção seguinte ficará aparente que padrões truncados tenham vaga semelhança com certos tipos de gráficos bimodais. Os dois geralmente podem ser distinguidos pela presença de pontos de inflexão em curvas cumulativas bimodais e a ausência de tais pontos de inflexão em curvas de distribuições truncadas.

- Como um exemplo de uma população lognormal simples consideramos os dados de produção de 74 depósitos de veios de Pb-Zn-Ag no campo mineiro de Ainsworth, sul de British Columbia. Estes dados foram compilados por Orr (1971) para todos os depósitos conhecidos no campo que tinham produzido uma tonelada ou mais de minério.

- Neste exemplo é conveniente plotar os logaritmos dos dados de produção sobre um papel de probabilidade logarítmica. A razão é que os dados abrangem mais que 5 ordens de magnitude e seria incômodo usar papel de log probabilidade geralmente usado.

- Estimativas de parâmetros da distribuição podem ser lidos diretamente do gráfico usando como base a linha ajustada aos pontos e lendo as ordenadas que correspondem aos percentis 16, 50 e 84.

- Comumente os dados plotados não formam uma linha reta sobre um papel de probabilidade mas têm uma curvatura definida com um ponto de inflexão pronunciado, ou mudam a direção de curvatura. Tais padrões comumente resultam da presença de duas (ou mais) populações em um grupo de dados. Atenção será dirigida primeiro a padrões ideais baseados em populações hipotéticas conhecidas, a partir das quais a atenção será dada para fazer generalizações que ajudem no processo reverso de extrair populações constituintes de populações na realidade misturadas. Procedimentos que são concernentes com a estimativa de populações constituintes de uma combinação constituintes de uma combinação de duas ou mais distribuições de densidade são conhecidos como partição (*partitioning*).

- Bölviken (1971) descreveu recentemente um método gráfico simples de determinar a curva de probabilidade para duas populações cujos parâmetros são especificados, combinados em quaisquer proporções desejadas. Consideramos duas populações A e B que estão combinadas em uma curva de probabilidade simples nas proporções 0,3A e 0,7B. Em qualquer nível de ordenada a % cumulativa das populações é igual a 0,3 vezes a % cumulativa de A mais 0,7 vezes a % cumulativa de B. A relação pode ser generalizada e expressa na forma de equação como segue:

$$P (A + B) = f_A . P_A + f_B . P_B$$

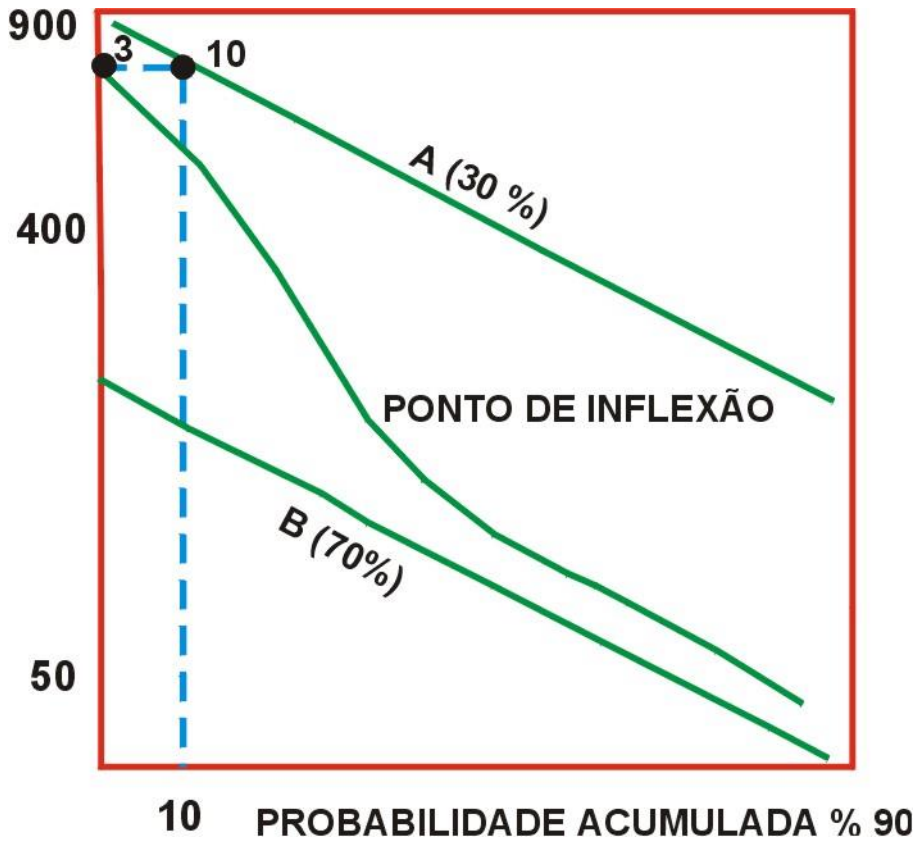
onde P (A+B) é a probabilidade cumulativa das populações combinadas, P_A a probabilidade cumulativa da população A, P_B a probabilidade cumulativa da população B, f_A a fração dos dados totais representados pela população A e f_B a fração dos dados totais representados pela população B. Devido essas duas populações compreenderem 100% dos dados, $(f_A + f_B) = 1$.

- Como exemplo, se no nível de ordenada 200 as curvas A e B são intersectadas em 78% e 2,1%, aplicando a equação teremos:

$$P (A + B) = 0,3 (78) + 0,7 (2,1) = 24,9\%$$

Assim, um ponto para as populações combinadas ao nível de ordenada 200 é calculado como sendo o percentual cumulativo 24,9.

- Partição de curvas de probabilidade bimodais não-intersectantes: o problema mais importante é certificar-se das proporções nas quais as duas populações estão presentes; um ponto de inflexão ocorre na curva bimodal precisamente no percentil cumulativo que representa as proporções das duas populações constituintes. Consequentemente, em se tratando com uma curva de probabilidade bimodal, a 1ª informação a ser tirada é o percentil cumulativo no qual ocorre o ponto de inflexão.



- Na figura acima, um ponto de inflexão é aparente ou perto do percentil cumulativo 30, indicando 30% de uma população superior e 70% de uma população inferior. O segmento central inclinado indica que há superposição de intervalos efetiva apreciável das duas populações (se não houvesse superposição o segmento central seria quase vertical). A extremidade superior da curva, entretanto, reflete somente a população superior e a extremidade inferior somente a população inferior. Consequentemente estas duas partes da curva podem ser usadas para estimar as populações constituintes, isto é, para repartir a mistura.

- Agora consideramos o percentil cumulativo em qualquer nível de ordenada perto do extremo superior da curva, como o percentil 3 para facilidade de cálculo. Este ponto representa somente 3% dos dados totais mas é (3/30) x 100 = 10% da população superior.

- A população inferior pode ser estimada pelo mesmo procedimento fornecendo a escala de probabilidade a leitura por uma forma complementar (por exemplo, o percentil cumulativo 80 é lido como 100-80= percentual cumulativo 20).

- O método é rápido mesmo com um mínimo de experiência mas os resultados devem sempre ser checados, particularmente no intervalo intermediário de superposição das duas populações onde os *trends* para ambas as populações repartidas foram extrapolados.

- Em casos práticos alguma dificuldade é comum na definição de um ponto de inflexão dentro de vários pontos de % cumulativa. Se um valor incorreto é escolhido, misturas ideais das populações repartidas não concordarão com a curva original e uma 2ª tentativa com um novo ponto de inflexão deve ser tentada. Tais tentativas seriam repetidas até que concordância aceitável seja obtida ou até que as populações-modelo (normal ou lognormal, conforme o caso) sejam rejeitadas.

- Devido à frequência com que gráficos de probabilidade bimodais não-intersectantes são encontrados na natureza, vale a pena sumarizar seus caracteres:

1) Tais gráficos descrevem uma curva contínua com forma característica, incluindo um segmento central mergulhante flanqueado por segmentos terminais menos mergulhantes;

2) Um ponto de inflexão ocorre no segmento central, a um % que define as proporções relativas das duas populações constituintes;

3) Alguma indicação dos desvios-padrão relativos de populações constituintes é dado por inclinações relativas dos 2 segmentos terminais fornecendo as duas populações não são muito diferentes em suas proporções;

4) Se as duas populações têm intervalo de superposição sem significância, o segmento central da curva é vertical. Quando a quantidade superposta aumenta a inclinação do segmento central se torna progressivamente menor.

5) O procedimento de partição inclui os seguintes estágios em seqüência:

   a) fazer uma curva suave através dos pontos plotados

   b) escolher o ponto de inflexão

   c) repartir tanto a população superior como a inferior

   d) escolher misturas ideais de população repartidas com a curva original descrevendo os dados reais

- Partição de curvas de probabilidade bimodais intersectantes é geralmente um procedimento mais difícil que o caso de curvas não-interserctantes e, como regra, envolve consideravelmente mais tentativa-e-erro. Há vários caracteres de curvas intersectantes ideais, entretanto, que ajudam na determinação do procedimento de partição mais eficiente.

1) As duas curvas para populações constituintes intersectam uma a outra e a curva bimodal no segmento central relativamente plano.

2) Esta intersecção tripla ocorre em um ponto de intersecção na curva bimodal. Infelizmente este ponto de inflexão é geralmente difícil, se não impossível, tirar precisamente, particularmente em dados reais para cujos pontos estão espalhados ao redor de uma curva generalizada.

3) O intervalo de ordenada abrangido pelo segmento central relativamente plano, razoavelmente acurado, o intervalo efetivo da pequena população dispersa.

4) O intervalo de % abrangido pelo segmento central fornece uma estimativa muito grosseira da proporção da pequena população dispersa representada na curva bimodal. Esta estimativa é geralmente alta, mas ao menos fornece um ponto inicial para partição por tentativa-e-erro. Pelo mesmo sinal, a estimativa complementar para a proporção da larga população dispersa é geralmente baixa.

- Quando todos os 3 segmentos de uma curva bimodal intersectante são bem definidos, a partição não é muito difícil. Um procedimento é o que se segue:

1) Uma estimativa da proporção da extensão da população dispersa é feita como em 4) acima.

2) Usando esta estimativa, pontos próximos ao terminal dos 2 segmentos finais são recalculados e plotados com uma população simples. Se os pontos recalculados plotados se alinham a proporção estimada é correta e a linha define a variação da amplitude da população. Se os pontos não se alinham uma nova proporção (geralmente maior) deve ser usada com base para novos cálculos. Este procedimento é repetido até que um padrão linear é obtido para pontos calculados representando a população partida.

3) Uma vez que a amplitude da população esteja definida, uma série de pontos sobre a população menor pode ser calculada usando a relação

$$P_C = f_W \cdot P_W + f_S \cdot P_S$$

onde, para qualquer ordenada, P_C é a probabilidade cumulativa das populações combinadas (isto é, a curva dos dados reais), P_W e P_S são % cumulativas das populações largamente e fracamente dispersas respectivamente, e f_W e f_S são frações (proporções) das populações largamente e reduzidamente dispersas respectivamente. Neste ponto no procedimento de partição P_S é o único não conhecido na equação.

4) A população menor dispersa é então estimada como uma linha reta feita através dos pontos determinados nos passos anteriores.

5) O procedimento de partição seria então checado com recalculagem de combinações ideais das populações partidas em vários níveis de ordenadas por comparação com a curva bimodal para dados reais.

Gráficos de probabilidades de combinações de 3 ou mais populações

- Em geral, crescendo o número de populações representadas em uma curva de probabilidade simples, aumenta a incerteza na interpretação da curva. Isto é particularmente verdadeiro nos casos de número de dados relativamente escasso. Para compensar esta dificuldade um procedimento útil é agrupar os dados com base em alguma propriedade física ou química fundamental, como tipo de rocha, pH, etc. para produzir curvas para uma interpretação simples. Em certos casos, entretanto, este agrupamento não é possível e uma curva de probabilidade polimodal deve formar a base de qualquer interpretação tentada. Como regra, interpretação de distribuições contendo somente 3 populações é direta, embora ambigüidade possa surgir.

- Consideremos 3 populações lognormais, A, B, e C, assumindo que estas populações estejam combinadas na proporção 1:1:1. É conveniente aqui combinar estas populações em 2 estágios sucessivos. Em 1º, as populações A e B são misturadas na proporção 1:1 para obter a curva (A+B). Então (A+B) e C são misturados na proporção 2:1 para dar a curva (A+B+C). Um exame da curva (A+B+C) mostra que 2 pontos de inflexão estão presentes. Como regra geral, o número de pontos de inflexão é um menos que o número de populações presentes. Além disso, como com o caso bimodal, os pontos de inflexão fornecem estimativas acuradas das proporções nas quais as 3 populações estão presentes, isto é, nos percentis cumulativos 33 e 67. Estas relações formam a base para partição de curvas de probabilidade de dados reais.

- Normalmente se 3 populações estão presentes e não se sobrepõem muito extensivamente sua partição é um procedimento direto não como o de uma curva bimodal, mas envolvendo um nível adicional. Considerável ambigüidade pode estar presente, entretanto, se relativamente poucos dados são disponíveis, ou se populações intersectantes estão presentes.

- Em certas curvas de probabilidade o procedimento de partição é simples. As populações maior e menor podem sofrer a partição exatamente do mesmo modo que as curvas bimodais. Com experiência é possível, em alguns casos, fazer a partição da população intermediária diretamente. Mais comumente, entretanto, é necessário usar um procedimento mais simples mas mais longo que envolve:

1) Combinação das populações alta e baixa nas proporções em que elas estão presentes (derivados de pontos de inflexão)

2) Aplicação da fórmula

$$P\ (A + B + C) = f_B\ .\ P_B + f(_{A+C)}\ .\ P_{(A+C)}$$

para vários níveis. Nesta equação P(A+B+C) e P(A+C) são lidos do gráfico de probabilidade. As 3 populações assim “partidas” seriam recombinadas em vários níveis ordenados para comparação com a curva dos dados reais.

Grupamento efetivo de dados polimodais estimativa de *thresholds* em dados geoquímicos

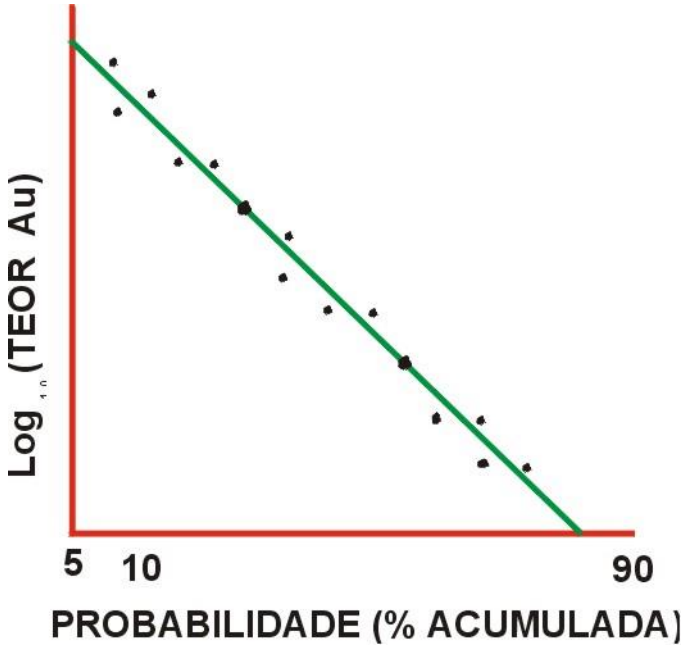
- Não há obviamente nenhuma razão por que somente 2,5% de qualquer grupo de dados necessite ser anômalo – por que não 1% ou 25%? Em um caso onde nenhuma superposição efetiva das duas populações ocorra, é relativamente

fácil escolher um *threshold* por rápido exame mesmo de um histograma ou dados tabulados (desde que a lista dos dados não seja muito longa).

- Entretanto, se as duas populações se sobrepõem mais e mais, a escolha de um *threshold* se torna bem mais difícil. Quanto mais, mais a efetividade de um valor simples de *threshold* diminui. Se ele é escolhido de tal modo que todos os valores anômalos sejam incluídos, então uma alta proporção de valores de *background* também são incluídos. Se ele é escolhido em algum ponto dentro do intervalo de superposição então um número valores anômalos são perdidos devido à inclusão com valores de *background*. Obviamente um procedimento é desejável para a escolha de valores do *threshold* que maximiza o reconhecimento de valores anômalos e minimiza o número de valores de *background* incluídos com valores anômalos. Gráficos de probabilidade cumulativa fornecem uma técnica gráfica efetiva para resolver este problema (Sinclair, 1974).

Gráficos cumulativos de dados consistindo de um pequeno número de valores

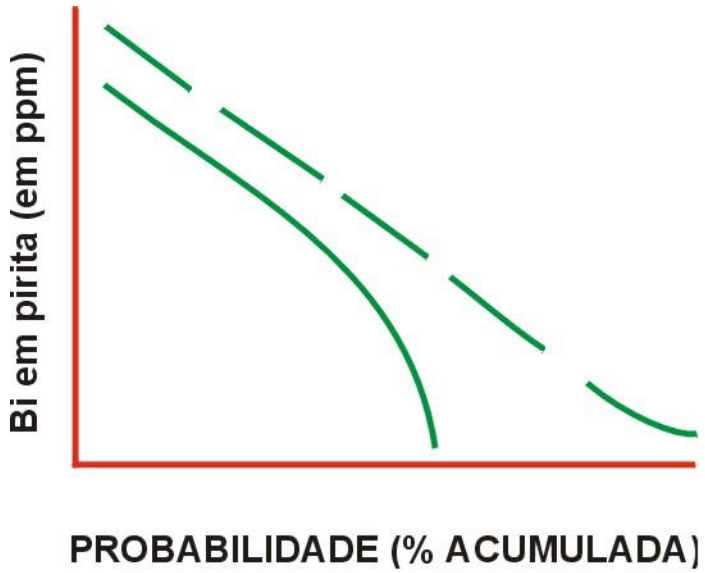
- Os dados não podem ser usados para constituir um gráfico cumulativo do modo antes explicado se eles incluem somente um papel de probabilidade, descrito em certos textos estatísticos (Schmitt, 1969) envolve as freqüências cumulativas de valores individuais em vez de freqüências cumulativas de intervalos. Como exemplo extremo desta técnica consideremos a média de ouro das análises de produção total de cada um dos 19 depósitos de veios em Ainsworth Mining Camp, Britsh Columbia. Note que neste caso logaritmos de valores de análise são plotados em papel de probabilidade aritmética, porque os valores aritméticos cobrem 4 ordens de magnitude e um gráfico de log-probabilidade teria sido incômodo usando papel gráfico disponível comercialmente. O gráfico deu uma reta:



- As inclinações deste gráfico para considerações exploratórias são dignas de menção. De 74 depósitos produtivos na área, somente 19 têm conteúdo de Au registrado em fontes publicamente disponíveis. Esses 19 valores então representam a amostra total sobre as quais a população toda pode ser estimada, e esta estimativa é um constituinte importante de qualquer minério potencial que possa ser achado no campo.

Dados reais contendo uma alta proporção de valores zero ou menores que o limite de detecção

- 67 amostras de pirita da mina de Mo de Endako e vizinhanças foram analisadas espectrograficamente para 12 elementos menores, incluindo Bi. Destas, em 26 não se detectaram valores de Bi. O gráfico para Bi:



- A curva é relativamente bem definida e tem a aparência de uma população simples plotada sobre parte do intervalo de probabilidade, nesse caso os 60% superiores do intervalo de probabilidade. Para checar se esta curva pode ou não representar uma população simples, um número de pontos sobre a curva suavizada foi recalculado a 100%. Está mostrado na mesma figura que se pode, com esses dados, traçar uma reta sobre os novos pontos sem dificuldade.

- Uma conclusão destas análises é que duas populações estão representadas nos dados – uma população lognormal superior, com 60% dos valores, acima do limite de detecção e quase perfeitamente definida pelo gráfico de probabilidade. Uma 2ª população consistindo de 40% dos dados, inferior ao

limite de detecção é coincidentemente um eficiente *threshold* separando as duas populações. É importante enfatizar que a curvatura do gráfico de probabilidade neste exemplo, baseada no total de 67 amostras, é o indicativo da existência de duas populações. Se houver curvatura, tentar estimar a proporção da população para a qual os dados são úteis, mesmo por reconhecimento de um ponto de inflexão, ou por tentativa e erro.

Plotando dados com falhas ou intervalos pequenos

- Há duas maneiras de tratar os dados que em muitos casos dão resultados idênticos. O 1º, recomendado por Lepeltier (1969), é escolher um intervalo suficientemente largo para que nenhum intervalo de um histograma normal fique vazio. O 2º método envolve valores individuais acumulados.

- Note as duas linhas da figura, a mais baixa das quais estima sem muita precisão a média devido ser baseada num intervalo muito estreito relativo à precisão do método analítico usado. Este problema existe com muitos dados analíticos que têm uma pequena variação relativa à precisão, e é particularmente comum com dados químicos semi-quantitativos onde valores não são interpolados entre padrões. Exemplos geoquímicos comuns incluem Mo, Hg e Ag.