ANÁLISE DE COMPONENTES PRINCIPAIS: FUNDAMENTOS DE UMA TÉCNICA DE ANÁLISE DE DADOS MULTIVARIADA APLICÁVEL A DADOS GEOLÓGICOS

José Leonardo Silva Andriotti[1]

ABSTRACT - Principal Component Analysis (PCA) is a multivariate technique by which a number of related variables is transformed in a new set of uncorrelated variables. PCA obtains linear transformations of the original variables since certain optimal conditions are achieved. Factor analysis (FA) is a technique similar in nature to PCA, but they have some important differences between them. PCA explains variability, FA explains correlations; FA has several estimation procedures, and the estimates are not unique. PCA has only one estimation procedure and its solution is unique. FA methodology usually includes a rotation to simple structure. As the results of PCA and FA are usually similar, it is a common practice to use PCA, since it is easier to use. However, each method optimizes different things. Confusions between PCA an FA are common in bibliography. Basic definitions and some examples of application of PCA and FA are presented and discussed in this text.

RESUMO - A Análise de Componentes Principais é uma técnica de análise de dados com crescente utilização em Geologia nos últimos anos. No presente trabalho é discutida, na parte inicial, a análise de dados multivariada como um todo, discorrendo-se sobre os seus princípios básicos e os objetivos a que se propõe. A Análise de Componentes Principais propriamente dita é tratada com mais detalhes, tanto no que diz respeito à sua conceituação teórica formal quanto no que toca às suas aplicações em Geologia, sendo listados vários trabalhos em que esta técnica tem sido aplicada a dados geológicos.

INTRODUÇÃO - Este documento visa à apresentação, de forma suscinta, dos fundamentos da técnica de Análise de Componentes Principais(ACP), tanto no que diz respeito aos seus fundamentos teóricos e de formulação matemática como de suas aplicações às ciências geológicas. A Análise de Fatores(AF), por suas diversas similaridades com a ACP, e mesmo devido à confusão reinante na bibliografia geológica sobre o uso da ACP e da AF, também é enfocada no presente texto. As

1 - CPRM – Rua Banco da Província, 105, Porto Alegre – RS – CEP 90840-035

razões que levam à utilização da técnica de ACP aos dados geológicos são expostas e discutidas, bem como é apresentada uma extensa lista de referências bibliográficas, específicas de Estatística Multivariada e aplicadas à Geologia especificamente, que foram consultadas e que abarcam os conceitos e as aplicações aqui discutidas. Esta técnica tem sido, em tempos recentes, cada vez mais utilizada nas diversas áreas que compõem a Geologia, e com resultados considerados pelos seus usuários como altamente satisfatórios.

ANÁLISE DE DADOS MULTIVARIADA

A Análise de Dados Multivariada é aquela que se ocupa de conjuntos de dados oriundos de diversas medidas obtidas sobre uma mesma amostra, e leva em conta, na busca de seus resultados, não somente as características individuais de cada uma das medidas obtidas sobre uma dada amostra mas, também, as relações porventura existentes entre as diversas variáveis utilizadas na investigação. Pressupõe, pois, a existência de diversas observações e, nelas, de diversas variáveis.

Se se levar em conta que as diversas variáveis interagem entre si para formar o quadro resultante de uma determinada situação sob investigação, e que por esta razão foram todas mensuradas, não há, em princípio, nenhuma razão para que o estudo se processe tão somente sobre cada variável individualmente. O estudo, ou o enfoque multivariado, se faz, então, imperativo.

De acordo com Kendall (1980), em um estudo multivariado vários enfoques podem ser utilizados, dentre os quais os mais importantes seriam, resumidamente, os seguintes:

1) Simplificação da estrutura dos dados, ou seja, a busca de uma representação simplificada do intrincado universo de investigação; isto se pode obter através da transformação, linear ou não, de um conjunto de variáveis interdependentes em um outro conjunto de dados independentes ou de menor dimensão;

2) Classificação, ou seja, a colocação das diversas observações em agrupamentos por similaridade, podendo-se agrupar, também, variáveis;

3) Análise de interdependência, incluindo desde independência total entre as diversas variáveis até a colinearidade, quando uma delas é combinação linear de algumas outras, ou, de modo mais amplo, quando uma delas seja uma função quantificável de outras;

4) Análise de dependência, pela qual se selecionam uma ou mais variáveis do conjunto original e se estuda a(s) sua(s) dependência(s) das demais, tal como ocorre nas técnicas de Regressão Múltipla ou na Análise de Correlação Canônica;

5) Formulação de hipóteses, encontrando-se modelos que permitam formular hipóteses em função de parâmetros estimáveis.

Em uma análise univariada quase sempre é possível obter-se uma boa caracterização da distribuição de probabilidades a partir de tão somente dois parâmetros, quais sejam a média aritmética e a variância. A inferência estatística exige que se obtenha uma amostra aleatória e, sobre ela, se calcule os melhores estimadores destes dois parâmetros, procedendo-se, a seguir, à interpretação das estimações efetuadas.

No caso multivariado se disporá, para o caso de haver n variáveis disponíveis, de n médias aritméticas, de n variâncias e de n(n-1)/2 covariâncias, que deverão ser estimadas e interpretadas.

Para a mente humana pode ser um pouco difícil trabalhar e raciocinar em um espaço multidimensional, pois que costumamos lidar com duas ou três dimensões. A maneira mais fácil de se aproximar desta noção é através da análise multivariada.

As técnicas multivariadas são muito potentes e de vasta utilização nos mais variados campos de investigação científica.

Entre os métodos de análise multivariada de dados que se concentram no estudo da interdependência entre variáveis e entre indivíduos citam-se a Análise de Fatores(AF), a Análise de Clusters, a Análise de Correlação Canônica e a Análise de Componentes Principais, e dentre os métodos para detecção de dependência estão a Análise de Regressão e a Análise Discriminante.

A Análise de Componentes Principais é uma das mais utilizadas, é uma técnica matemática que não exige que a distribuição de probabilidades da população sob estudo seja conhecida, não necessitando, pois, de um modelo estatístico que explique a estrutura probabilística dos erros, ressaltando-se que, em sendo possível assumir-se uma distribuição multinormal para os dados disponíveis se pode encontrar significação estatística nas componentes, pois que poder-se-á associar a cada uma delas uma medida de confiabilidade.

ANÁLISE DE COMPONENTES PRINCIPAIS

Karl Pearson, em 1901, publicou um trabalho sobre o ajuste de um sistema de pontos em um espaço multivariado a uma linha ou a um plano, trabalho este que foi retomado por Hottelling em 1933. Este trabalho de Hottelling contém a definição da técnica de Análise de Componentes Principais (ACP) tal como é conhecida até nossos dias.

Hottelling concentrou seus estudos nas componentes que tinham a maior variabilidade do sistema de pontos, ou seja, que respondiam pela maior parte da variabilidade dos dados sob estudo, daí o nome Principal acoplado ao próprio nome da técnica.

A Análise de Componentes Principais é uma técnica, sob o enfoque de Hottelling, que permite se encontre uma forma de classificar os pontos e/ou de detectar relações entre eles. Assim, a Análise de Componentes Principais deve ser usada quando houver a necessidade de se conhecer a relação entre os elementos de uma população e haja a suspeita de que, na citada relação, haja a influência de um conjunto de variáveis ou de propriedades dos elementos disponíveis.

A Análise de Componentes Principais encontra um campo fértil para sua utilização quando se tem em mãos um conjunto de dados multivariados, isto é, quando houver muitas variáveis interagindo concomitantemente no fenômeno ou no processo estudado, e que não se possa postular, com base nos dados disponíveis, uma estrutura particular destas variáveis. Esta situação é comum em Geologia.

Os trabalhos de Johnson e Wichern(1982), de Chatfield e Collins(1980), de Bouroche e Saporta(1980) e, muito especialmente, de Jackson(1991) apresentam em detalhe os fundamentos indispensáveis para um bom conhecimento das técnicas de ACP e AF. São obras específicas de Estatística Multivariada, sendo que a última é específica sobre a Análise de Componentes Principais e mais recente, dando uma idéia muito clara sobre as características do método, suas vantagens e desvantagens.

O trabalho de Cibois(1983) é específico sobre a Análise de Fatores(ou Análise Fatorial), sendo igualmente muito elucidativo sobre a técnica tratada.

Guillaume(1977) e McCammon(1975) são obras específicas que abordam a utilização das várias técnicas estatísticas multivariadas em Geologia, com ênfase à Prospecção Geoquímica, e em Le Maître(1982) são apresentadas aplicações específicas à Petrologia.

Pla(1986) também trata especificamente da técnica de Análise de Componentes Principais, desde seus fundamentos teóricos até suas aplicações às várias áreas onde possa ser aplicada, incluindo exemplos de aplicação numérica.

Quando se tem conhecimento da existência de uma ou de várias variáveis independentes, e, portanto, de outro conjunto de variáveis dependentes, pode-se aplicar as técnicas de regressão múltipla ou as de regressão multivariada.

Se se tiver conhecimento prévio da inexistência de relações entre as variáveis, ou seja, que haja independência ou ausência de correlação entre as elas, deve-se abster-se de procurar uma explicação para a "relação" entre as variáveis, ou mesmo entre os indivíduos a partir destas variáveis. Se tal ocorrer (ausência de correlação entre as diversas variáveis disponíveis), deve-se estudar as variáveis individualmente à luz das técnicas univariadas, as quais nos darão resultados igualmente consistentes e ao mesmo tempo menos complicados.

A transformação que gera as Componentes Principais pode ser vista como uma operação que gera um novo conjunto de dados, ou um novo conjunto de coordenadas que sejam perpendiculares entre si.

As Componentes Principais são, então, não correlacionadas entre si, ao contrário do que ocorre com os dados originais (as variáveis realmente medidas: em Prospecção Geoquímica, os teores dos elementos, por exemplo).

No caso da Prospecção Geoquímica as diversas variáveis podem apresentar algum tipo de relação entre si, relação própria da composição química das litologias a que pertencem ou, mesmo, de alguma característica própria da área trabalhada. As Componentes Principais obtidas não necessariamente refletirão relações, pois que representarão associações litológicas, processos geológicos que atuaram na área ou, mesmo, eventos carreadores de mineralizações, que podem ter sido gerados em diferentes épocas geológicas ou sob diferentes condicionantes geológicas.

Não vem ao caso discutir-se o quão complicado possa ser um método qualquer de análise de dados, em especial quando se trabalha com conjuntos de dados multivariados. As assim chamadas "complicações" algébricas embutidas em um método que manipula grandes quantidades de variáveis e de observações são evidentes, mas os recursos computacionais hoje disponíveis tornam de importância secundária esta preocupação: tanto a nível de software quanto a nível de hardware as capacitações crescem continuamente.

O problema capital de qualquer técnica é a interpretação dos seus resultados, a associação correta dos valores numéricos obtidos à Ciência à qual pertencem os dados utilizados.

Em Geologia é de fundamental importância a obtenção da relação entre as Componentes Principais obtidas e fatos ocorridos na área geográfica à qual pertencem os dados, ou seja, as Componentes devem responder a processos plausíveis de ocorrerem naquela situação geológica, naquelas suítes presentes na região, no condicionamento estrutural a que foi submetida a área.

Muitas vezes estes condicionamentos, estes processos, não estão muito claros, requerem um refinamento interpretativo maior, pois que estão disfarçados ou mesmo escondidos e ofuscados por outros acontecimentos de registro mais ostensivo na crosta terrestre; isto é muito comum em Geologia. As técnicas estatísticas multivariadas e as técnicas computacionais são um importante aporte à interpretação de dados geológicos, não substituem nem nos levam a prescindir de nenhuma das ferramentas a que estamos habituados a utilizar quando em presença de um problema geológico.

A Análise de Componentes Principais é sensível às magnitudes das medidas das variáveis tratadas. Assim, se tivermos uma variável cujos valores estejam expressos em ppm (partes por milhão) e outra que esteja expressa em ppb (partes por bilhão), esta última terá uma influência tão maior que a primeira quanto for o quociente entre suas unidades de medida (no caso, o resultado da divisão de um bilhão por um milhão), o que é completamente indesejável e, mesmo, funesto para a interpretação. Para se contornar este problema, entretanto, se dispõe de soluções, como, por exemplo, a padronização das variáveis(a padronização de uma variável é uma técnica que consiste em transformar a variável de modo a que ela passe a ter média aritmética igual a zero e variância igual à unidade, obtida através da divisão da diferença de cada dado com relação à sua média aritmética pelo desvio padrão da referida variável; os elementos da matriz variância-covariância consistirão, deste modo, de correlações, e as Componentes Principais serão produzidas em forma adimensional).

Algebricamente, se as variáveis forem todas medidas em mesmas unidades, as Componentes Principais são calculadas a partir da matriz de variância-covariância, ou matriz de dispersão; a matriz de correlação só é utilizada no caso citado acima, qual seja o de haver diferenças de magnitude nas unidades das diversas variáveis que entram no processo algébrico de solução do problema através da técnica de Análise de Componentes Principais.

A Análise de Componentes Principais determina os eixos principais de uma configuração multidimensional, bem como fornece as coordenadas de cada indivíduo da população relativamente a estes eixos. Isto tem a vantagem de que, com os resultados da Análise de Componentes Principais em mãos, os dados podem ser representados em poucas dimensões através de projeções dos diversos pontos ortogonalmente sobre os eixos principais. Deste modo, se os dados forem plotados através dos escores das suas componentes um e dois, ou seja, das duas Componentes Principais que representam as maiores variabilidades dos dados, o plano resultante será o melhor sumário bi-dimensional do espaço n-dimensional original, em que n é o número de variáveis originalmente disponíveis para estudo.

A Análise de Componentes Principais é, pois, uma técnica de análise multidimensional linear cujo objetivo é classificar os elementos de um conjunto (as nossas observações originais) em classes de elementos próximos ou similares (nuvens de elementos) e de estabelecer o balanço de correlações entre as variáveis originais utilizadas no estudo.

A Análise de Componentes Principais permite se responda às seguintes questões, dentre outras:

1) Qual é o ponto mais próximo da nuvem das observações? A nuvem está concentrada neste ponto?

2) Qual é o eixo mais próximo da nuvem das observações? A nuvem está concentrada sobre este eixo?

3) Qual é o plano mais próximo da nuvem das observações? A nuvem está concentrada sobre este plano?

4) Qual é o sub-espaço de dimensão K mais próximo da nuvem das observações? A nuvem está concentrada no interior deste sub-espaço?

A Análise de Componentes Principais busca transformar p variáveis originais correlacionadas entre si em q variáveis não correlacionadas entre si , $q < p$, isto é, em q Componentes Principais que sejam funções lineares das variáveis originais. Em outras palavras, é feita uma transformação do espaço dimensional do problema, transformação esta que se dá de tal modo que a projeção das observações originais no novo sistema seja nula (isto é, a projeção seja ortogonal). Deste modo, o eixo original (variável p) é transformado em um novo eixo (chamado Componente Principal), o qual é estatisticamente independente, pois que cada Componente Principal é, por sua vez, ortogonal à seguinte.

Quando da utilização desta técnica se estuda e analisa, também, a correlação entre as variáveis originais e as Componentes Principais; para tal se calcula todas as correlações de cada variável original com cada nova variável (cada Componente Principal gerada).

É importante assinalar que a ACP é uma técnica essencialmente descritiva e tem uma interpretação geométrica, desde a formulação de Karl Pearson, de 1901, em planos de melhor ajuste e vetores de máxima concentração, em função de distâncias euclidianas.

Um cuidado que se deve ter, sempre, é sobre a possibilidade da presença de dados aberrantes no nosso conjunto original de dados. Um valor aberrante denuncia sua presença por apresentar um valor, para uma dada variável, extremamente elevado ou extremamente baixo para a mesma na situação estudada. A presença de dados aberrantes não ocorre apenas no campo univariado: ela é, apenas, de mais fácil detecção quando ocorre nestes casos (estudos univariados).

Os dados aberrantes se manifestam não apenas como uma observação de valor exageradamente elevado ou baixo no caso multivariado, pois que não existe, num estudo multivariado, um modo único de se ordenar os dados que contêm múltiplas medições. Além disto, a presença de dados aberrantes em dados multidimensionais pode causar distorções não somente nas chamadas medidas de posição, como média aritmética, ou de dispersão, como o desvio padrão, mas igualmente pode introduzir distorções sobre as chamadas medidas de orientação, ou seja, sobre as correlações entre as diversas variáveis estudadas. Além disto, há que se considerar a existência de uma certa variedade de tipos de observações aberrantes, já que um vetor-resposta, ou uma observação (ou uma amostra, em sentido geológico) pode ser aberrante por conter um grande erro em um de seus valores ou, mesmo, por haver um erro sistemático em todos os valores numéricos que a compõem. Em estudos multivariados a situação é bem mais complexa, pois que uma observação pode ser aberrante para um determinado propósito sem sê-lo para outro obrigatoriamente.

A Análise de Componentes Principais é um método multivariado que se presta também, e de forma muito eficaz, para a identificação de valores aberrantes em um determinado conjunto de dados.

Os valores extremos das primeiras Componentes Principais (as que respondem pela maior parte da variabilidade dos dados originais) são úteis na identificação das observações que contribuem para aumentar grandemente a variância e a covariância (ou a correlação se for o caso, já citado anteriormente, de se usar a matriz de correlações ao se usar a técnica de ACP), e as últimas Componentes Principais são sensíveis para detectar observações que contenham dimensões consideradas anormais para os dados sob investigação.

Através da elaboração de diagramas de dispersão entre pares das primeiras e das últimas Componentes Principais se pode identificar visualmente as observações aberrantes.

Além de ser possível plotar em gráficos as variáveis originais em função das Componentes Principais se pode, também, plotar os valores das primeiras Componentes Principais para cada observação.

A Análise de Componentes Principais (ACP) tem como um de seus objetivos a transformação de p variáveis originais em q variáveis não correlacionadas, sendo $q < p$, em que as q variáveis geradas são chamadas Componentes Principais e são funções lineares das p variáveis originais. É realizada uma transformação do espaço dimensional de tal sorte que a projeção das observações originais no novo sistema seja nula, ou seja, projeção ortogonal; assim, o eixo original (Variável Original) é transformado no novo eixo (Componente Principal), o qual é estatisticamente independente, pois que cada Componente Principal, por sua vez, é ortogonal à seguinte.

FORMULAÇÃO MATEMÁTICA

A ACP determina os eixos principais de uma configuração multidimensional e as coordenadas de cada indivíduo em relação a estes eixos.

Cada nova variável (Componente Principal, CP) gerada é uma combinação linear das p variáveis originais, podendo ser escrita como

$$CP = A_{11}\, p_1 + A_{12}\, p_2 + \ldots\ldots + A_{ij}\, p_j$$

em que p1, p2, ... pj são as variáveis originais e em que A_{11}, A_{12}, ... A_{ij} são coeficientes (referidos na literatura como latent roots, characteristic roots, proper values ou eigenvalues, ou ainda, em língua portuguesa, como autovalores). Estes coeficientes são tais que satisfazem:

1) Variância CP_1 > Variância CP_2 > ..... > Variância CP_q,

2) os valores de quaisquer dois CPs são não correlacionados, o que não ocorre com as variáveis originais, e
3) para qualquer CP a soma dos quadrados destes coeficientes vale um.

A variância ao longo do primeiro eixo principal é maior que a variância ao longo de qualquer outra linha reta possível de ser traçada através dos pontos, sendo seguida em magnitude pela variância ao longo do segundo eixo, e assim sucessivamente. Em outras palavras: as novas variáveis (as CP) são geradas em ordem decrescente de importância, de modo que a primeira CP responda pela maior parte da variação total observada nos dados originais, seguida pela segunda CP, terceira CP, etc.

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

A Análise de Fatores (ou Análise Fatorial) é muitas vezes confundida com a ACP; são, entretanto, conceitualmente distintas.

A Análise Fatorial (AF) pode ser considerada como uma extensão da ACP.

Na ACP é buscada a solução com máxima variância de todas as variáveis disponíveis, em AF um número pré-determinado de fatores (também menor que o número de variáveis originais) é definido para responder maximamente pela inter-correlação entre as variáveis observadas. Cada CP maximiza a variância comum, ou seja, a ACP é orientada para a variância, e a AF é orientada para a correlação. A pré-determinação do número de fatores na AF é ponto de muita controvérsia. Não é raro encontrar-se, na bibliografia, aplicações de uma destas técnicas sendo referida como a outra.

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

A plotagem dos escores das duas primeiras componentes para cada indivíduo é uma maneira útil de encontrar clusters nos dados.

As Componentes Principais são geralmente modificadas pelas escalas e elas não são uma característica única dos dados. Se, por exemplo, uma variável tiver uma variância muito maior do que todas as outras então esta variável dominará a primeira Componente Principal da matriz de covariância qualquer que seja a estrutura de correlação, enquanto que se as variáveis estiverem todas em uma escala para terem variância unitária, então a primeira Componente Principal será muito diferente em tipo. Por causa disto geralmente se pensa ser de pouco benefício executar ACP a menos que as variáveis tenham variâncias "grosseiramente similares", o que pode ser o caso, por exemplo, se todas as variáveis forem percentagens, ou tiverem sido medidas com as mesmas coordenadas.

A maneira convencional de contornar o problema de escalas é analisar a matriz de correlação em vez de analisar a matriz de covariância.

Todas as variáveis sendo padronizadas para ter variância igual a um faz com que, de alguma forma, tenham igual importância. Se as variáveis forem pensadas não terem a mesma importância, então a análise da matriz de correlação não é recomendada. Analisar a matriz de correlação também dificulta a comparação dos resultados da ACP de duas ou mais amostras diferentes.

Os objetivos principais da ACP seriam a identificação de novas variáveis significativas subjacentes, a redução da dimensionalidade do problema e a eliminação das variáveis originais que contribuem relativamente pouco em termos de informação extra.

O procedimento de usar os escores das componentes para produzir um mapa dos indivíduos é algumas vezes chamado de Análise de Coordenadas Principais ou Classical Scaling.

ACP pode revelar, como dito acima, clusters de variáveis que não seriam encontrados por outros meios.

Uma alternativa utilizada é a execução de regressão, não sobre as variáveis originais, mas sobre as primeiras (poucas) Componentes Principais.

ACP consiste em uma transformação ortogonal no espaço de dimensão p, das variáveis originais, em um novo conjunto de variáveis, chamadas Componentes Principais. Os estágios principais da análise são:

a) decidir se se deve incluir todas as variáveis originais, e se algumas das variáveis precisam ser transformadas

b) calcular a matriz de correlação ou de covariância, tendo em mente que um coeficiente de correlação geralmente não seria calculado para um par de variáveis cuja relação fosse obviamente não linear

c) verificar a matriz de correlação e observar quaisquer grupos naturais de variáveis que tenham "altas" correlações. Se todas as correlações forem "pequenas", entretanto, então provavelmente não se terá vantagem em proceder à ACP

d) calcular os eigenvalues(autovalores) e eigenvectors(autovetores) da matriz de correlação ou de covariância

e) examinar os eigenvalues e tentar decidir quantos são "grandes". Isto indicaria a dimensionalidade efetiva dos dados

f) verificar os grupos de variáveis sugeridos pelas componentes e considerar se as componentes têm alguma interpretação significativa

g) usar os escores das componentes em análises subseqüentes como uma forma de reduzir a dimensionalidade do problema.

Poderemos, entretanto, recair sobre as dificuldades da técnica, a saber:

i) pode ser difícil ou perigoso tentar ler significados demasiados nas componentes

ii) as componentes não são invariantes sob transformações lineares das variáveis

iii) não há nenhum modelo estatístico subjacente, e nenhum aprovisionamento pode ser feito para as componentes da variância devidas ao "erro". Isto significa que o comportamento da amostragem dos eigenvalues e eigenvectors é desconhecido. Um resultado é que não há uma maneira objetiva de decidir quantos eigenvalues são "grandes". Outra conseqüência é que é difícil comparar diferentes componentes resultantes da execução de ACP sobre duas ou mais amostras diferentes do mesmo tipo.

O uso de ACP ou AF em Exploração Geoquímica tem sido geralmente para separar associações de elementos inerentes na estrutura da matriz de correlação em um número de grupos de elementos que, juntos, respondem pela maior parte da variabilidade observada dos dados originais.

Técnicas que tomam a matriz de correlação como ponto de partida são a grosso modo referidas como Modo R.

Por exemplo, se nós tivermos resultados analíticos de Cu e Zn em um conjunto de amostras de sedimentos de corrente, podemos aventurar a opinião de que o capeamento de Mn sobre os sedimentos causa sua coprecipitação e isto responde por sua variância comum.

A variância total, ou a soma das variâncias de cada variável, pode ser pensada como sendo composta de duas partes: a variância comum (C^2) que é comum a todas as variáveis, e as variâncias específicas (S_i^2) para cada uma das variáveis individuais. A variância total pode então ser dividida em uma soma de C^2 e S_1^2, S_2^2,....., S_p^2. Técnicas multivariadas existem para responder por ou para fazer uso de C^2. A variância comum é um termo aritmético que representa a porção da variância total que é devida a combinações de 2,3....., k das variáveis do conjunto. A porção remanescente é formada de uma variação específica a cada uma das variáveis separadamente.

De acordo com Jöreskog et al. (1976) o contraste é que em ACP cada componente é determinada de forma a maximizar C^2 (isto é, para responder pela máxima variância de todas as variáveis observadas) enquanto que em Análise de Fatores(AF) um número dado de "fatores" (menor que o número de variáveis) é definido de forma a responder maximamente pela intercorrelação entre as variáveis observadas. Assim, cada componente na ACP pode ser dita maximizar a variância comum. A ACP é, então, orientada para a variância, enquanto que a AF é orientada para a correlação.

Para dados geoquímicos seria sugerido que a ACP é favorável em situações em que a amplitude de variação dos elementos é característica do ambiente geoquímico, enquanto que a AF é favorável em situações em que associações de elementos caracterizam o ambiente geoquímico.

Na matriz de covariância as variáveis com valores iniciais elevados tenderão a dominar aquelas com pequenos valores; podemos aplicar ponderações subjetivas quando no início se trabalha com as escalas das variáveis (por exemplo % Fe x 100) ou usar alguma outra transformação como logaritmo, raiz quadrada ou transformação do tipo power. Obviamente, no cálculo da matriz de correlação de Pearson os dados são transformados para terem igual peso pela padronização no curso do cálculo. A matriz de correlação é geralmente utilizada em trabalhos geoquímicos para prevenir influências indevidas dos grandes valores sobre os resultados.

Uma matriz não singular, simétrica, pxp, como a matriz de covariância S, pode ser reduzida a uma matriz diagonal L através da operação com uma matriz ortonormal U tal que

$$U'SU = L$$

Os elementos da diagonal de L, l_1, l_2,...., l_p são as raízes características, raízes latentes ou eigenvalores de S.

As colunas de U, μ_1, μ_2,...., μ_p são os vetores característicos ou eigenvetores de S.

As raízes características são obtidas da equação característica

$$|S - \ell I| = 0$$

onde I é a matriz identidade.

O ponto de partida da ACP é a matriz de covariância S

$$S = \begin{vmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ s_{2p} & \cdots & \cdots & s_p^2 \end{vmatrix}$$

Se as covariâncias não forem zero significa haver relação linear entre as variáveis. Os eixos coordenados das novas variáveis são descritos pelos vetores característicos μ_i que formam a matriz U usados na transformação

$$z = U'[\, x - \bar{x} \,]$$

onde x e $\bar{x}$ são os px1 vetores de observações sobre as variáveis originais e suas médias.

As variáveis transformadas são chamadas Componentes Principais de x, e as observações individuais transformadas são chamadas escores z.

1) o determinante da matriz de covariância |S| é chamado variância generalizada, sua raiz quadrada é proporcional à área ou ao volume gerado pelos dados;

2) a soma das variâncias das variáveis

$$s_1^2 + s_2^2 + \ldots\ldots s_p^2 = \mathrm{Tr}(S) = \text{traço de S}$$

$$|S| = |L|\, \ell_1 \ell_2 \ldots \ell_p$$

isto é, o determinante da matriz de covariância original é igual ao produto das raízes características, ou seja, a soma das variâncias originais é igual à soma das raízes características , que são as variâncias das Componentes Principais. A razão de cada raiz característica em relação ao total indica a proporção da variabilidade total explicada por cada CP. Uma vez que as raízes características são estimativas amostrais, estas proporções também o são.

É possível também determinar a correlação de cada CP com cada variável original; a correlação da i-ésima CP (z_i) com a j-ésima variável original (x_j) é igual a

$$r_{zx} = \frac{\mu_{ji}\sqrt{l_i}}{s_j}$$

A 1ª CP é mais altamente correlacionada com as variáveis originais que a 2ª, o que é esperado porque ela explica mais variabilidade que a 2ª. A soma dos quadrados de cada linha vale um.

Cada variável é uma combinação das CP.

Ao se falar de população, a matriz de covariância é Σ, e as raízes características serão, $\lambda_1, \lambda_2, \ldots\ldots, \lambda_p$.

Os vetores característicos re-escalonados, os vetores U, são ortogonais e têm comprimento unitário. Usá-los resulta em CPs não correlacionadas e de variâncias correspondendo às raízes características; são os vetores re-escalonados à unidade.

ANÁLISE DE FATORES

O objetivo essencial da AF é descrever, se possível, as relações de covariância entre muitas variáveis em termos de umas poucas subjacentes, mas não observáveis, quantidades aleatórias chamadas Fatores. Basicamente o modelo fatorial é motivado pelo seguinte argumento: suponhamos que variáveis possam ser agrupadas pelas suas correlações. Todas as variáveis dentro de um grupo particular são altamente correlacionadas entre si mas têm correlações relativamente pequenas com variáveis em um grupo diferente. É concebível que cada grupo de variáveis represente uma construção subjacente simples, ou Fator, que é responsável pelas correlações observadas.

A AF pode ser considerada como uma extensão da ACP. Ambas podem ser vistas como tentativas para aproximar a matriz de covariância. Entretanto, a aproximação baseada no modelo de AF é mais elaborada. A questão primária em AF é saber se os dados são consistentes com uma estrutura pré-determinada.

AF é um nome geral para uma variedade de procedimentos para examinar as correlações dentro de um conjunto de dados.

Muitos autores têm alargado o termo para incluir técnicas como ACP que, diferentemente da AF, não assume nenhum modelo matemático bem definido. Como resultado, pode-se encontrar exemplos na literatura geológica que alegadamente tratam com AF quando na realidade apenas ACP foi usada e vice-versa. Há diferenças distintas entre os conceitos e as exigências da ACP, que é orientada para a variância, e a AF, que é orientada para a covariância ou para a correlação.

As aplicações geológicas da AF são geralmente confinadas a dois métodos; um é chamado AF Modo R e investiga as relações entre variáveis, enquanto o outro investiga as relações entre objetos e se chama AF Modo Q.

Enquanto alguns autores questionam se a AF Modo R trouxe alguma coisa nova em relação à ACP, uma extensão do método Modo Q provou ser uma técnica útil em Petrologia.

APLICAÇÕES EM GEOLOGIA

No item **REFERÊNCIAS BIBLIOGRÁFICAS** são listados quarenta e dois artigos técnicos com aplicações das técnicas de ACP e AF a dados geológicos nos mais variados lugares do mundo, todos eles trazendo alguma informação valiosa para o bom entendimento destas técnicas.

Dentre eles, entretanto, citaremos alguns que julgamos merecerem um destaque por razões que acompanham cada uma das citações em separado.

Ajayi(1981), Morsy(1993) e Bellehumeur, Marcotte e Jébrak(1994) se destacam por apresentarem aplicações destas técnicas à Prospecção Geoquímica por sedimentos de corrente, apresentando seus resultados juntamente com fundamentos das técnicas e justificando de forma clara os motivos das opções que adotaram.

La Roche e Isnard(1978) apresentam uma aplicação destas técnicas à litogeoquímica; Andriotti, Nardi e Bitencourt(1989) da mesma forma, utilizando dados de uma área situada no Estado do Rio Grande do Sul.

Em termos de aplicações a dados de solos vale destacar o trabalho de Barbier e Wilhelm(1978), que apresentam de forma resumida as conclusões a que chegaram pela aplicação destas técnicas a nada menos que dezessete estudos de casos franceses.

Van Andel(1964) é um dos trabalhos clássicos de aplicação no apoio à interpretação de resultados obtidos através de campanhas de amostragem de minerais pesados; outro trabalho muito importante nesta área, bem mais recente, é o de Bellehumeur e Jébrak(1993).

Mackiewicz e Ratajczak(1993), na revista Computers & Geosciences, apresentam um enfoque dos aspectos computacionais relacionados à Análise de Componentes Principais; a revista na qual publicaram seu trabalho é específica em aplicações de informática às ciências geológicas.

Jimenez-Espinosa, Souza e Chica-Olmo(1993) apresentam uma conjugação de aplicações das técnicas de Análise de Componentes Principais e de Krigagem Fatorial a dados geoquímicos, aliando aplicações de Estatística Multivariada e de Geoestatística ao mesmo conjunto de dados.

A maior quantidade de trabalhos específicos a que se teve acesso apresenta aplicações a dados obtidos através de campanhas de sedimentos de corrente, seguido pelos trabalhos de aplicação a dados de litogeoquímica; apenas alguns poucos artigos apresentam aplicações a malhas de solos.

CONCLUSÕES

As técnicas de Análise de Componentes Principais e de Análise de Fatores são de larga utilização em Geociências, sendo seus resultados considerados pelos vários autores que as utilizaram como tendo sido revestidos de sucesso.

Elas apresentam utilidade tanto no estudo específico de áreas comprovadamente mineralizadas(estudos de detalhe) quanto quando aplicadas a grandes áreas, visando a dar apoio à caracterização do fundo geoquímico regional.

Suas aplicações têm crescido muito nos anos recentes, o que pode ser comprovado ao se consultar diversas publicações técnicas específicas das ciências geológicas, com destaque à Geoquímica e à Petrologia.

No Rio Grande do Sul, entretanto, a utilização deste enfoque a dados geológicos, tal como tem ocorrido em países como África do Sul, Austrália, Canadá e Estados Unidos, entre outros, e mesmo em outros estados brasileiros(os anais de diversos eventos estão plenos destas aplicações), se é que tem ocorrido não tem se refletido em publicações, ou seja, não tem sido divulgada para a comunidade geológica.

Por ser o enfoque multivariado o único meio de tratamento de dados que leva em conta as interações entre as diversas variáveis, além de considerar e respeitar as suas características individuais, julga-se imprescindível a sua utilização quando na presença de dados de natureza multivariada: tais são os dados geoquímicos obtidos em campanhas de sedimentos de corrente, de rochas, de solos e de concentrados de minerais pesados, na esmagadora maioria dos casos de que se tem conhecimento.

Julgamos, como conclusão, que é imperativo se utilize, para qualquer caso e para qualquer situação com que nos defrontemos, as técnicas adequadas, desde que disponíveis. Este é o caso das técnicas de Análise de Componentes Principais e de Análise de Fatores a dados geológicos, quando estivermos na presença de conjuntos de dados multivariados.

REFERÊNCIAS BIBLIOGRÁFICAS

As referências bibliográficas abaixo listadas formam uma lista de livros-texto e de artigos técnicos específicos de aplicação das técnicas de Análise de Componentes Principais e de Análise de Fatores em Geologia. Nestes textos se pode encontrar uma vasta gama de aplicações práticas das técnicas aqui descritas nas ciências geológicas.

AJAYI, T. R. - 1981 - Statistical analysis of stream sediment data from the Ife - Ilesha area of SW Nigeria. Journ. of Geoch. Expl., vol. 15, p. 539 - 548

ANDRIOTTI, J.L.S.; NARDI, L.V.S. e BITENCOURT, M. F. - 1989 - Aplicação de Análise de Fatores ao Estudo Geoquímico do Complexo Granítico de Caçapava do Sul. IV Simp. Sul-Brasileiro de Geologia, P. Alegre, RS

BARBIER, J. e WILHELM, E. - 1978 - Superficial geochemical dispersion around sulphide deposits: some examples in France. Journ. Geoch. Expl., vol. 10, p. 1 - 39

BELLEHUMEUR, C. e JÉBRAK, M. - 1993 - Regional heavy mineral survey in the exploration for gold using regression: Grenville Province, southwestern Quebec. Journ. of Geoch. Expl., vol. 47, p. 45 - 61

BELLEHUMEUR, C.; MARCOTTE, D. e JÉBRAK, M. - 1994 - Multi-element relationships and spatial structures of regional geochemical data from stream sediments, SW Quebec, Canada. Journ. of Geoch. Expl., vol. 51, p. 11 - 35

BELLIDO, F. e BRÄNDLE, J. L. - 1979 (Abril) - An application of Q - Mode Factor Analysis to the geochemical study of a granitic pluton (La Cabrera, Sistema Central, Spain) - Sci. la Terre, Série Informatique Géologique, n° 13, p. 111 - 123

BETTENCOURT, J. S. e LANDIM, P. M. B. - 1974 - Estudo geoquímico de óxidos e elementos traços de rochas calcárias do Grupo Açungui pela Análise Fatorial. XXVIII Congr. Bras. Geol., vol. 7, p. 153 - 160

BOUROCHE, J. - M. e SAPORTA, G. -1980 - L'Analyse des Données. Presses Universitaires de France, 127 p.

CHAPMAN, R. P. - 1978 - Evaluation of some statistical methods of interpreting multi-element geochemical drainage data from New Brunswick. Math. Geol., vol. 10, n° 2, p. 195 - 224

CHATFIELD, C. & COLLINS, C. - 1980 - Introduction to Multivariate Analysis. Chapman and Hall, N.Y., 246 p.

CHORK, C. Y. e SALMINEN, R. - 1993 - Interpreting exploration geochemical data from Outokumpu, Finland: a MVE - robust factor analysis. Journ. of Geoch. Expl., vol. 48, p. 1 - 20

CIBOIS, P. - 1983 - L'Analyse Factorielle. Presses Universitaires de France, 127 p.

DE VIVO, B. et al. - 1993 - Detailed Geochemical survey in the Peloritani Arc (NE Sicily, Italy): evidence of gold anomalies. Journ. of Geoch. Expl., vol. 46, p. 309-324

ELUEZE, A. A. e OLADE, M. A. - 1985 - Interpretation through factor analysis of stream-sediment reconnaissance data for gold exploration in Ilesha greenstone belt, southwest Nigeria. Trans. Inst. Min. Metall. p. B155 - B160

FILIPPINI , J. M. - 1991 - Aplicação de AF em Prospecção Geoquímica: amostragem de sedimentos de corrente em terrenos Pré-Cambrianos do Uruguai. 3° Congresso Brasileiro de Geoquímica, São Paulo, p. 546 - 550

GARRETT, R. G. e NICHOL, I. - 1969 - Factor Analysis as an aid in the interpretation of regional geochemical stream sediment data - Colorado School of Mines, Quarterly, vol. 64, n° 1, p. 245 -264

GRIPP, A. H. - 1986 - Análise Fatorial de Correspondências - aplicação ao estudo de dados de prospecção geoquímica. XXXIV Congr. Bras. Geol., vol. 6, p. 2842 - 2849

GUILLAUME, A. - 1977 - Introduction à la Géologie Quantitative. Masson, Paris, cap.VI, 200 pp

HESP, W. R. e RIGBY, D. - 1975 - Aspects of tin metallogenesis in the Tasman Geosyncline, Eastern Australia, as reflected by Cluster and Factor Analyses. Journ. of Geoch. Expl. , 4, p. 331 - 347

HOLLAND, P.; BEATY, D. & SNOW, G. - 1988 - Comparitive Elemental and Oxygen Isotope Geochemistry of Jasperoid in the Northern Great Basin: Evidence for Distinctive Fluid Evolution in Gold - Producing Hydrotermal Systems. Econ. Geol., vol. 83, p. 1401 - 1423

HOTTELLING, H. - 1933 - Analysis of a Complex of Statistical Variables into Principal Components. J. Educ. Psychol., 24, p.417 - 4451 e 498 - 520

HOWART,R.J. - 1983 - Statistics and Data Analysis in Geochemical Prospecting. Elsevier, N.Y., capítulo 6 (Howart,R.J. E Sinding-Larsen,R. - Multivariate Analysis, p.207 - 289)

IMBRIE, J. e VAN ANDEL, Tj. H. - 1964 - Vector Analysis of heavy - mineral data. Geol. Soc. of Am. Bull., vol. 75, p. 1131 - 1156

JACKSON, J. E. - 1991 - A User's Guide to Principal Components. John Wiley & Sons, N.Y., 569 p.

JIMENEZ - ESPINOSA, R.; SOUSA, A. J. e CHICA-OLMO, M. - 1993 - Identification of geochemical anomalies using PCA and factorial kriging analysis. Journ. of Geoch. Expl., vol. 46, p. 245-256

JOHNSON,R.A. e WICHERN,D.W. - 1982 - Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, 594p.

JÖRESKOG, K. G. et al. - 1976 - Geological Factor Analysis. Elsevier Pub., N. Y., 177p.

KELLEY, K. D. e KELLEY, D. L. - 1992 - Reconnaissance exploration geochemistry in the central Brooks Range, northern Alaska: implications for exploration of sediment - hosted Zn - Pb - Ag deposits. Journ. of Geoch. Expl., vol. 45, p. 273 - 300

KENDALL, M. G. - 1980 - Multivariate Analysis. Griffin, Londres, 2ª ed., 210p.

KLOVAN, J. E. - 1966 - The use of Factor Analysis in determining depositional environments from grain-size distributions. Journ. of Sedim. Petrol., vol. 36, n° 1, p. 115 - 125

LA ROCHE, H. de e ISNARD, P. - 1978 (nov.) - A comparison between "conventional" and "statistical" processing of 330 rock analyses on a regular network sampling in an association of biotite granites and two mica leucogranites. Sciences de la Terre, Série Informatique Géologique, n° 12, p. 65 - 98

LE MAÎTRE, R. W. - 1982 - Numerical Petrology. Elsevier Sci. Pub., cap. 7 e 8, 281p.

LITAOR, M. I. e KOYUMDJISKY, H. - 1989 - Factor Analysis of a Lithosequence in the Northestern Samaria Steppe (Israel). Geoderma, 44, p. 1 - 15

MACEDO, A. B. e RÜEGG, N. R. - 1974 - Avaliação de elementos traços nas rochas basálticas da Bacia do Paraná estudada por meio da Análise Fatorial. XXVIII Congr. Bras. Geol., vol. 7, p. 49 - 56

MACEDO, A. B. e RÜEGG, N. R. - 1974 - Aplicação da Análise Fatorial ao estudo de elementos principais nas rochas basálticas da Bacia do Paraná. XXVIII Congr. Bras. Geol., vol. 7, p. 143 - 151

MACKIEWICZ, A. e RATAJCZAK, W. - 1993 - Principal Components Analysis (PCA). Computers & Geosciences, vol. 19, n° 3, p. 303 - 342

MATALAS, N. C. e REIHER, B. J. - 1967 - Some Comments on the Use of Factor Analyses - Water Persounces Research, vol. 3, n° 1, p. 213 - 223

McCAMMON, R.B. - 1975 - Concepts in Geostatistics. Springer-Verlag, N.Y., cap.2, 168pp.

MEDEIROS NETO, F. A. - 1987 - Características geoquímicas dos sedimentos de corrrente em diversos ambientes geológicos da Folha de Carajás, SE do Pará - Geochimica Brasiliensis, v. 1, n° 2, p. 235 - 246

MORSY, M. A. - 1993 - An example of application of FA on geochemical stream sediment survey in Umm Khariga arla, Eastern Desert, Egypt. Math. Geol., vol. 25, n° 7, p. 833 - 850

OLIVEIRA, S. M. B.; BRUNS, R. e LOPES, L. M. - 1988 - Análise Fatorial Varimax de dados geoquímicos de materiais lateríticos. Aplicações à prospecção geoquímica. XXXV Cong. Bras. Geol., vol. 4, p. 1765 - 1780

RONDINELLI, D. et al. - 1989 - Geoquímica Regional da Folha Pilar do Sul - São Paulo. II Congr. Bras. Geoquímica, RJ, p. 47 - 60

SAAGER, R. e ESSELAAR, P. A. - 1969 - Factor Analysis of geochemical data from the Basal Reef, Orange Free State Goldfield, South Africa. Econ. Geol., vol. 64, p. 445 - 451

SAAGER, R. e SINCLAIR, A. J. - 1974 - Factor Analysis of stream sediment geochemical data from the Mount Nansen Area, Yukon Territory, Canada. Mineral. Deposita, 9, p. 243 - 252

SAHA, A. K.; SARKAR, S. N.; BASU, S. e GANGULY, D. - 1986 - A multivariate statistical study of Cooper mineralization in the Central Section of Mosaboni Mine, Eastern Singhbhum, India. Math. Geology, vol. 18, n° 2, p. 215 - 235

SAMAMA, J.C.; ROYER, J.J. e N'GANZI, C. - 1989 - Prise en Compte de la Surface Spécifique des Prévèlements en Prospection Géochimique: exemple de l'uranium dans les sédiments de ruisseau. Journ. of Geoch. Expl., 32, p. 453 - 466

SELINUS, O. - 1981 - Lithogeochemical exploration data in sulphide prospecting in northern Sweden. Journ. of Geoch. Expl., vol. 15, p. 181 - 201

TILL, R. e COLLEY, H. - 1973 - Thoughts on use of PCA in petrogenetic problems. Math. Geol., vol. 5, n° 4, p. 341 - 350

TRIPATHI, V. S. - 1979 - Factor Analysis in Geochemical Exploration. Journ. of Geoch. Expl.,v. 11, p. 263 - 275

USUNOFF, E. J. e GUZMÁN - GUZMÁN, A. - 1989 - Multivariate Analysis en Hydrogeoclemistng: on example of the ese of factor and correspondence Analyses. Ground Water, vol. 27, n° 1, p. 27-34

VAN DE HAAR, A. J. e VAN GAANS, P. F. M. - 1993 - Hydrothermal alteration of the Beira Schists around the W - Sn specialised Regoufe granite, NW Portugal. Journ. of Geoch. Expl., vol. 46, p. 335 - 347

VOGT, J. H. e KOLLENBERG, W. - 1987 - Genetic implications of geochemical factor analysis in a sediment-hosted Cu-Pb-Ba mineralization. Mineral. Deposita, 22, p. 151 - 160

WEBER, L. e DAVIS, J. C. - 1990 - Multivariate statistical analysis of stream sediment geochemistry in the Grazer Paläozoikum, Austria. Min. Depos., p. 213 - 220

YAMAMOTO, J. K.; GOULART, E. P. e HASUI, Y. - 1980 - Análise estatística dos dados químicos do Complexo Alcalino de Anitápolis, SC. XXXI Congr. Bras. Geol., v. 2, p. 1272 - 1283