· U C ·
FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Aplicação multivariavel na caracterização de sedimentos de corrente

**Dissertação apresentada para provas de Mestrado em Química,
Ramo de Controlo de Qualidade e Ambiente**

António Carlos Tavares dos Santos

Mestrado em Química

Departamento de Química

FCTUC

Agosto 2010

UNIVERSIDADE DE COIMBRA

Aplicação multivariavel na caracterização de sedimentos de corrente

António Carlos Tavares dos Santos

Dissertação apresentada para provas de Mestrado em Química, ramo de Controlo de Qualidade e Ambiente

Orientador: Professor Doutor Jorge Costa Pereira

Co-orientador: Professora Doutora Maria Manuela da Vinha Guerreiro Silva

Agosto de 2010

Universidade de Coimbra

Agradecimentos

Este trabalho representa o fim de um ciclo e o início de outro. No ciclo que agora se fecha, muitas foram as pessoas que, de uma forma ou de outra, contribuíram para que esta caminhada fosse menos cansativa, com palavras de incentivo, com conversas e ideias que me ajudaram a ver a vida de uma forma diferente. A todas elas agradeço e dedico-lhes este trabalho.

Agradeço à minha família por estar sempre a meu lado, por me dar palavras de conforto e, por vezes, conselhos sábios que me ajudam a ultrapassar as situações difíceis do dia-a-dia.

Aos meus amigos, por me apoiarem nos momentos mais chatos e quando estou mais em baixo.

Ao Doutor Jorge Costa Pereira, por ter aceite ser meu orientador e, desta forma, me ensinar a usar, de uma forma mais adequada e proveitosa, as ferramentas analíticas que se encontram ao nosso dispor e que, nem sempre as usamos ou porque não sabemos que elas existem, ou então porque não sabemos como usa-las. Agradeço-lhe, ainda, por ter acreditado nas minhas capacidades para a realização deste trabalho, encorajando-me e dando-me ânimo para continuar e não desistir.

À Doutora Manuela da Vinha que, desde que a convidei para co-orientadora, se mostrou disponível e interessada na ideia, estando sempre pronta para tirar as dúvidas que surgissem. Agradeço-lhe, ainda, pela críticas construtivas que me fez, pois permitiram elaborar este trabalho de uma forma concisa e de fácil compreensão.

À Professora Doutora Ana Margarida Neiva, por me ter dado algumas facilidades em termos de horários de trabalho.

À Doutora Paula Carvalho pela cedência dos dados que são a base deste estudo, e, ainda, aos demais elementos do Grupo de Geoquímica do Centro de Geociências pela companhia e pelo esclarecimento de dúvidas pontuais sobre questões ligadas à geoquímica.

Agradeço ainda aos funcionários do Departamento de Ciências da Terra com quem tenho convivido ao longo deste último ano, a companhia e os momentos de alegre distracção que ajuda a passar alguns momentos de maior desalento.

Objectivos

Neste trabalho pretende-se atingir os seguintes objectivos:

- perceber até que ponto uma análise quimiométrica com métodos robustos pode ajudar a interpretar dados geoquímicos;

- aferir e tentar incrementar procedimentos e algoritmos de análise multivariada de dados;

- demonstrar que, recorrendo a métodos mais robustos, como o PCA ou o PLS, se consegue extrair mais informação dos dados a analisar.

Resumo

Este trabalho tem como objectivo mostrar que, para além dos métodos estatísticos convencionais (média, desvio padrão, variância, teste t-student, entre outros), também se pode recorrer a métodos mais robustos, como o PCA, o LDA ou o PLS e, desta forma, conseguir extrair mais informação dos dados a analisar.

Assim, foram utilizadas amostras de sedimentos de corrente, recolhidas no Distrito de Castelo Branco (em Portugal), na zona de Sarzedas, uma vez que estes são reconhecidos como sendo um meio privilegiado para a propagação de focos de poluição dos solos. Depois de sujeitas ao plano de amostragem, inicialmente estabelecido, a partir de métodos reconhecidos internacionalmente para a recolha e preparação deste tipo de amostras, sofreram digestão ácida parcial, sendo analisadas por ICP-OES, e, nos casos em que o limite de detecção do ICP-OES não é suficientemente preciso, colorimetria.

Após a análise, obteve-se uma matriz de dados que continha os valores das concentrações de cada variável, que foi devidamente tratada, removendo-se as amostras que não continham alguns dos valores, evitando, assim, complicações aquando da aplicação dos métodos estatísticos.

A análise dos dados revelou que apenas a distribuição da variável bário (Ba) pode ser considerada normal, que existem variáveis que não contêm informação relevante para o estudo realizado, uma vez que se mantêm constantes e, finalmente, que há variáveis que possuem baixa capacidade discriminante e severos casos de outliers, confirmados depois de usado o teste t-student.

Depois de removidos os outliers, procedeu-se ao seu escalamento e posterior análise robusta, a qual demonstrou que com apenas 13 das 26 variáveis se consegue explicar 80% da informação contida na matriz de dados original e que existem várias associações, entre as variáveis, com significado estatístico e visual (V-Fe, V-Ba, Fe-Ba, V-Cu, V-Ni, Fe-Cu, Fe-Ni, Fe-Zn, Ni-Cu, Fe-Cr e V-Cr).

Abstract

This work aims at showing that, in addition to conventional statistical methods (mean, standard deviation, variance, t-student test, among others), we can use more robust methods such as PCA, LDA or PLS to be able to extract much information

Thus, we used samples of stream sediments collected in the district of Castelo Branco (Portugal), area of Sarzedas, since these are recognized as a privileged means for the spread of outbreaks of soil pollution. After being placed under the sampling plan, originally adopted from internationally recognized methods for the collection and preparation of such samples, suffered partial acid digestion and they were analyzed by ICP-OES, and where the accuracy of ICP-OES was low, colorimetry.

After analysis, we obtained a data matriz containing the concentration values for each variable. This matriz has been properly treated by removing the samples that did not contain some values, thus avoiding complications in the application of statistical methods.

Thus, the conventional statistical methods revealed that only the distribution of the barium (Ba) can be considered normal and that there are variables that do not contain information relevant to the study that is done, since they remain constant. You can also verify that there are variables that have low discriminating capacity and severe outlier cases confirmed after using the t-student test.

Once removed these outliers, we proceeded to its escalation and further robust analysis of which, being noticeable that only 13 of the 26 variables that can explain 80% of the information contained in the original data matriz, revealed that there are several associations variables, with statistical significance and visual (V-Fe, V-Ba, Ba-Fe, V-Cu, V-Ni, Fe-Cu, Fe-Ni, Fe-Zn, Ni-Cu, Fe-Cr and Cr-V).

Prefácio

Antes de mais convêm deixar aqui uma breve explicação acerca do tema sugerido e do caminho percorrido, nesta fase de formação profissionalizante.

Origem

O tema sugerido para o presente trabalho resultou da minha actividade profissional, já que, no dia-a-dia estou envolvido na análise de amostras de solo e sedimentos de corrente provenientes de diversas partes do País.

Sendo formado em Química e estando a frequentar o Mestrado em Química, na vertente de Controlo da Qualidade e Ambiente, sentimos que este assunto é deveras importante e se encontra perfeitamente enquadrado neste plano de estudos.

Contudo, esta actividade laboral, além de estar sujeita a sigilo (amostras confidenciais) não apresenta um cariz bem direccionado, destinado a um determinado fim específico, o que dificultou a progressão do trabalho.

Por outro lado, pelo Laboratório Químico, do Departamento de Ciências da Terra, da Universidade de Coimbra, têm passado diversos trabalhos cuja análise de resultados está limitada no tempo escasso que cada formando dispõe para prosseguir na sua aprendizagem.

Por sugestão do orientador interno, Doutor Jorge Costa Pereira, seria interessante repescar algum conjunto significativo de dados (base de dados) para testar e procurar desenvolver procedimentos e algoritmos de análise multivariada de dados, cada vez mais necessária no dia-a-dia.

Ainda por sugestão da orientadora externa, Professora Doutora Manuela da Vinha, e com o apoio da Professora Catedrática do Departamento de Ciências da Terra, Professora Doutora Ana Margarida Neiva, achou-se pertinente utilizar, para este fim, o conjunto de dados obtido pelo Instituto Geológico e Mineiro, no âmbito de campanha de prospecção geoquímica, pelo antigo serviço de fomento mineiro, compilado entre 1916 e 1951 e da posse da Doutora Paula de Carvalho. Trata-se de uma tabela extensa em que diversos sedimentos recolhidos na região de Sarzedas (distrito de Castelo Branco), são compilados com base em 26 variáveis (parâmetros de análise química), perfazendo um total de 684 amostras.

Ao tempo desta investigação, década de 50, apenas se concluiu que, nessa zona havia quantidades assinaláveis de Volfrâmio, Ouro e Estanho, tendo este estudo estado assim na origem de subsequentes explorações mineiras.

Estrutura

A presente tese encontra-se estruturada em cinco capítulos essenciais, a saber, introdução, fundamentação, procedimentos, resultados e discussão.

Na introdução apenas se lançam algumas ideias fundamentais e explicativas iniciais. Na fundamentação são desenvolvidos os assuntos que estão patentes e servem de suporte à parte experimental e discussão de resultados. Na parte de procedimentos procurou-se, de uma forma simplificada e não fastidiosa, dar orientações mais ou menos concretas, no sentido de se poder reconstruir e desenvolver o trabalho. Já nos resultados e discussão são, sempre que possível, apresentados resultados e explicações sobre as constatações observadas. Na parte de conclusões pretende-se deixar, em resumo, as contribuições mais relevantes deste trabalho.

Simplificações

O tratamento multivariado de dados exige a utilização de software adequado que segue a notação Americana (Internacional) para a representação de números reais.

Assim sendo, utilizou-se a notação Internacional para representar os valores reais em que o ponto representa o separador decimal, em vez da notação portuguesa. Esta simplificação permitiu, de uma forma expedita e coerente, transferir a informação de cálculo obtida com folhas de cálculo e com outras ferramentas de cálculo, para o texto sem outro tipo de dificuldades de conversão de formato.

A informação numérica em formato científico surge, também, por razões óbvias, sob a forma exponencial adequada às folhas de cálculo. Sempre que possível e adequado, os valores estimados estão representados com a sua incerteza associada, respectivo erro padrão, indicada entre parêntesis, seguida das respectivas unidades, de forma a conferir maior significado estatístico aos resultados obtidos.

Abreviaturas e Símbolos

(σ ;s_x) – desvio padrão

$\bar{X}$ - valor esperado (média aritmética com n << 1)

SS - valor mínimo da hipersuperfície de erro

Δ - Desvio sistemático, tendência [bias]

μ - Valor esperado (média aritmética com $n \to \infty$)

Color. – Colorimetria

e - erro absoluto (desvio do resultado)

EAA – Espectroscopia de Absorção Atómica

Eh – Potencial electroquímico

equação – equação

ex-IGM – ex-Instituto Geológico e Mineiro

H0 - hipótese nula

H1 - hipótese alternativa

ICP-OES ou **ICP** – Espectroscopia de Emissão Óptica com plasma acoplado indutivamente (do Inglês, Inductively Coupled Plasma Optical Emission Spectrometry)

ISO - Organização Internacional de Normalização (em inglês International Organization for Standardization)

IUPAC – União Internacional de Química Pura e Aplicada (do Inglês International Union of Pure and Applied Chemistry)

IVs - Infravermelhos

KLT – Transformador de Karhunen-Loève

Kurt - curtose

LD – Limite de Detecção

LQ – Limite de Quantificação

OLS – Método de ajuste explícito de funções por mínimos quadrados ordinários (do Inglês, Ordinary Least Squares)

p[H0] - valor de prova (α)

p[Norm] – probabilidade de que a distribuição seja normal

PC – Componente Principal (do Inglês Principal Component)

PCA – Método de ajuste explícito de funções por análise de componentes principais (do Inglês, Principal Component Analysis)

PCR - Método de ajuste explícito de funções por Regressão de Componentes Principais (do Inglês Principal Components Regression)

PLS – Método de ajuste explícito de funções por mínimos quadrados parciais (do Inglês, Partial Least Squares)

POD – Método de ajuste explícito de funções por decomposição ortogonal própria (do Inglês Proper Orthogonal Decomposition)

r - variável escalada

s - desvio padrão (n << 50)

SEM – Modelagem de Equações Estrutural (do Inglês Structural Equation Modeling)

Skew - estimativa de simetria

Teste-F – Teste de Fisher

T_{SK} – teste combinado de simetria e curtose

TV – valor de teste

Var – variância

WLS – Método de ajuste explícito de funções por Mínimos Quadrados Ponderados (do Inglês Weighted Least Squares)

x - variável independente

Xf – Matriz de dados filtrada

Xm – média

Xmedian – mediana

y - variável dependente

z - variável normalizada

α - erro tipo I (falsa rejeição)

β - erro tipo II (falsa aceitação)

ε - erro aleatório

σ – desvio padrão (n →∞)

μL - micro-litros

Índice

Capítulo 1

Introdução

O solo é a camada superficial da Terra, substrato essencial para a biosfera terrestre, cuja principal função consiste em ser a base de toda a cadeia alimentar, suportando e fornecendo os nutrientes essenciais aos organismos produtores e não só [1], ao mesmo tempo que intervém na regularização do ciclo hidrológico, através das suas capacidades de transformação, servindo como filtro, tampão ou permutador de iões. Contudo, estas podem estar comprometidas devido aos constantes contaminantes e poluentes nele introduzidos, pelo Homem, na realização das suas tarefas diárias, que podem diminuir a sua faculdade de renovação. Este facto é deveras alarmante, já que é no solo que se situam muitos dos aquíferos que abastecem a maioria das populações com água potável.

Assim, o solo é um importante elemento físico, patrimonial e paisagístico para o desenvolvimento e sustentabilidade das actividades humanas, tendo já sido declarado internacionalmente [2], que é fundamental protege-lo e limitar os seus processos de degradação.

Os solos podem ser contaminados naturalmente pela meteorização e erosão das rochas que possuam elementos tóxicos com teores elevados, pela existência de mineralizações ou depósitos minerais, pelas erupções vulcânicas e pela libertação de gases e de hidrocarbonetos. Todavia, a maioria da contaminação e poluição deve-se à acção do homem.

De facto, de acordo com o Instituto Nacional dos Resíduos, em Portugal Continental, os problemas de solos contaminados estão directamente relacionados com um desenvolvimento industrial insustentável, abuso de fertilizantes e pesticidas, deposições atmosféricas resultantes das várias actividades poluidoras, lixeiras a céu aberto, despejos ilegais de material poluente (por exemplo, descargas de fábricas), o armazenamento de combustíveis e substâncias perigosas em locais ou condições impróprios e a actividade mineira [1].

As amostras colhidas e processadas têm como intuito, de algum modo, poder contribuir no sentido de procurar obter indicações físico-químicas deste problema na região de Sarzedas, em Castelo Branco, onde a extracção mineira tem causado um largo impacto ambiental.

Na verdade, por as explorações mineiras constituírem um grande foco de contaminação das áreas envolventes [3 e 4], estão sujeitas, na CEE, a um controlo ambiental apertado [5], mas mesmo as que se encontram abandonadas continuam a contaminar e a poluir os solos envolventes [6, 7 e 8]. Perante todos estes “ataques omnipresentes”, é imprescindível a realização de um inventário dos locais potencialmente contaminados, com o intuito de criar uma base de dados para futuras acções concertadas de detecção, identificação, remediação e descontaminação desses solos.

Segundo o primeiro Inventário Nacional sobre Solos Contaminados, feito pelo então Instituto

dos Resíduos, em 2000, concluiu-se que existiam cerca de 22 mil locais com solos e aquíferos contaminados devido às actividades predominantes. Todavia, ainda hoje, Portugal continua sem legislação nacional sobre este tema, restando apenas as normas seguidas em países mais avançados, como o Canadá, Holanda ou Itália.

A contaminação dos solos torna-se ainda mais preocupante, quando ameaça indivíduos e bens, acentuando-se com a transferência dos elementos poluentes, através, por exemplo, dos sedimentos de corrente ou sedimentos de linhas de água, que alargam a zona contaminada [9].

Os sedimentos de corrente formam-se a partir de diversos processos, de entre os quais destacamos o efeito do pH na água. Neste caso, ocorre uma corrosão das rochas, por acção da água que adquire pH ácido, ao entrar em contacto com a atmosfera, que resulta da dissolução do dióxido de carbono, originando ácido carbónico. Este ácido vai ser o responsável pela oxidação de rochas e minerais[1], formando iões que, mais tarde, a água se encarrega de transportar. Para além da variação do pH, a alteração do Eh[2], temperatura, concentração do oxigénio e concentração de iões, levam, igualmente, à precipitação e/ou dissolução, contribuindo, assim, para a composição dos sedimentos [9 e 10].

Foi efectuado um estudo quimiométrico, usando formulações matemáticas, mais propriamente a Análise de Componentes Principais (PCA) e o Cálculo dos Mínimos Quadrados (PLS). Para permitir a melhor caracterização de grupos e aumentar o poder descriminante dos objectos, foi, também, realizado um estudo de valores discrepantes e de variáveis redundantes. Foram analisadas amostras de sedimentos de corrente, em que as variáveis são elementos alcalino-terrosos, metais de transição, actinídios, semi-metais, metais presentativos e não – metais, sendo que alguns destes elementos constituem os elementos de traço, e outros são micro nutrientes, que são os nutrientes que as plantas usam em quantidades muito pequenas, mas cuja ausência pode limitar o seu crescimento, e até mesmo matá-las. Os mais importantes são magnésio, ferro, cálcio, entre outros [11]. Por sua vez, na geoquímica o elemento de traço corresponde a um elemento químico, cuja concentração é inferior a 1000 ppm ou 0,1 % na composição de uma rocha [12].

[1] Um mineral é qualquer substância natural, sólida inorgânica, em que cada um tem uma estrutura e uma composição química definidas, que lhe confere um conjunto único de propriedades físicas [13].
[2] Eh – Potencial electroquímico.

Capítulo 2

2.1 Sedimentos de Corrente

Desde há muito tempo que o Homem começou a tentar classificar os sedimentos com base na sua textura. Aliás, as próprias populações, de forma intuitiva, os classificam, aplicando terminologias que, com frequência, foram adoptadas pela comunidade científica, como "lodo", "argila", "areia", "cascalho", “seixo” e “balastro” [9].

Como já foi levemente enunciado na Introdução, os sedimentos de corrente são materiais naturais de composição variável, devido à diversidade de materiais rochosos dos quais derivam, às contaminações naturais e às alterações induzidas pelo homem nas águas e nos solos. São, deste modo, ao serem transportados e depositados pela água, um meio privilegiado para a transferência de contaminantes [3].

A composição dos sedimentos de corrente depende da constituição das rochas e dos solos existentes na área de drenagem e das condições de drenagem, assim como das reacções químicas que se processam na água e entre esta e os sedimentos [3]. Como tal, os sedimentos apresentam uma textura muito variável, e a sua composição tem fracções distintas. Na verdade, há dois tipos principais de fracções: a fracção mais grosseira (> 2 mm), que é a carga de fundo e se reconduz a fragmentos de rochas e minerais, transportados por deslocamento (deslize e rolamento); e a fracção mais fina (<2 mm), que é particularmente útil para se estimar o grau de contaminação, distinguindo, também, as fontes naturais das antrópicas[3] [11].

Nesta última categoria integra-se a fracção argilosa dos sedimentos, uma vez que juntamente com o material orgânico constituem a fracção principal do sedimento onde se concentram os transportadores geoquímicos activos (hidróxidos, óxidos, sulfuretos e carbonatos), obtendo, consequentemente, a maior concentração de metais pesados [11], tanto de origem natural quanto antrópica. Ora, esta concentração só é possível graças à capacidade de troca catiónica, facultada pela grande superfície específica dos minerais argilosos, como a clorite $[(Mg,Fe)_3(Si,Al)_4O_{10}(OH)_2 \cdot (Mg,Fe)_3(OH)_6]$, caulinite $[Al_2Si_2O_5(OH)_4]$, ilite $[(Na,K)Al_2(Si_3AlO_{10})(OH_2)]$, esmectite $[½(Ca,Na)Al_3MgSi_8O_{20}(OH)_4nH_2O$]e à facilidade de transporte das argilas pelas águas superficiais, por terem uma granulometria muito fina [12].

Com efeito, a distribuição não uniforme dos elementos nas diferentes fracções granulométricas, causa variações nos conteúdos dos metais nas amostras [14], sendo necessário corrigir as concentrações dos diversos contaminantes e as influências da variação natural, que se observa na composição dos sedimentos, de forma a comparar, correctamente, os resultados obtidos

[3] Fontes antrópicas – são aquelas que sofreram efeitos ou processos causados pelas actividades humanas.

em diferentes locais. Assim, é importante efectuar uma normalização física [15], isto é, eliminar o efeito do tamanho das partículas, através da separação de fracções granulométricas, com um método de clivagem com peneiros a diferentes granulometrias, de acordo com o método estabelecido no início do trabalho.

Na verdade, os elementos traço presentes nos sedimentos podem apresentar-se em diferentes formas químicas influenciadas pela existência, ou não, de actividade antrópica na sua formação. Por outro lado, quando os elementos químicos são incorporados, precisamente a partir de actividades como a antrópica, apresentam uma mobilidade maior, estando ligados a outras fases do sedimento, como carbonatos, óxidos, hidróxidos e sulfetos. O comportamento dos ambientes, em que os sedimentos se encontram, varia consoante a forma de criação dos sedimentos e os processos que controlam o transporte e redistribuição dos elementos, como adsorção, dessorção, precipitação, solubilização e floculação [16].

Estes sedimentos são, essencialmente, constituídos pelos fragmentos de rochas, minerais e solos que foram desagregados e removidos, ou seja erodidos, pela acção das águas de escorrência e sedimentados no leito das linhas de água quando esta perde a sua capacidade de transporte; isto é, em linhas gerais pode dizer-se que resultam da interacção contínua dos processos de meteorização e erosão [17].

Há algumas regras formuladas que são geralmente seguidas, relativamente à substituição isomórfica de um ião por outro:

- um elemento menor pode substituir extensivamente um maior se os raios iónicos não diferirem mais do que cerca de 15 %;
- iões cujas cargas diferem por uma unidade podem substituir-se um pelo outro, desde que os seus raios iónicos sejam similares e a diferença de carga possa ser compensada por outra substituição. Por exemplo, o feldspato plagioclásio Na^{+} facilmente substitui Ca^{2+}, e a diferença de carga é compensada pela substituição de Si^{4+} para Al^{+3};
- de dois iões, que podem ocupar a mesma posição numa estrutura de cristal, o que faz as ligações mais fortes com seus vizinhos é aquele com o menor raio, maior carga, ou ambos. a substituição de um ião por outro pode ser muito limitado, mesmo quando o critério de tamanho é cumprido, se os vínculos formados diferirem marcadamente no carácter covalente [18].

Na prática, o estudo da variabilidade geoquímica dos sedimentos de corrente, tem como finalidade a cartografia geoquímica, a determinação dos fundos geoquímicos regionais [19-21], a prospecção de depósitos minerais e a inventariação e estudo de locais contaminados [3,12 e 21-27]. Na verdade, uma amostra de sedimento é representativa dos solos e das rochas existentes a montante

do seu local de colheita, por isso a sua área de influência varia entre dezenas a centenas de quilómetros quadrados [19, 20 e 27].

De acordo com o referido no parágrafo anterior, as amostras de sedimentos a montante da zona de erosão são as mais indicadas para a análise do grau de poluição ambiental, atribuída aos elementos-traço ou a substâncias tóxicas orgânicas [28], essencialmente nas águas superficiais. Normalmente, é nestas águas que os contaminantes se encontram mais concentrados e adsorvidos na fracção sólida suspensa nos cursos de água [29]. Esta fracção suspensa na água sedimenta, contribuindo para a composição dos sedimentos, quando a energia do curso de água diminui ou quando variam algumas características físicas e químicas, como temperatura, pH, salinidade, potencial redox e teores de quelantes orgânicos na água, que até podem provocar a remobilização para a fase aquosa [29-32].

Deste modo, as relações entre a composição dos sedimentos de corrente e a composição da água permitem determinar a partição água/sedimento de contaminantes e os modos e meios de dispersão destes [33]. Neste sentido, e graças à elevada capacidade de retenção e acumulação de elementos traço contidos na coluna de água, os sedimentos de corrente são utilizados como indicadores ambientais [34], ajudando o seu estudo a analisar o impacto ambiental de explorações mineiras, lixeiras a céu aberto, despejo de esgotos ilegais das fábricas para os solos que, por lixiviação, vão parar aos rios, lagoas, lençóis freáticos e mais tarde ao mar [35]. Ou seja, os sedimentos de corrente permitem conjecturar sobre as condições a que o sistema esteve ou está sujeito, uma vez que, pode apresentar a capacidade de actuar como registo histórico das actividades desenvolvidas na bacia hidrográfica.

Contudo, também a variação das condições ambientais pode originar uma remobilização dos metais que se concentram nos sedimentos de um rio ou reservatório, permitindo a sua reentrada para a coluna de água [36]. Desta forma, os sedimentos de corrente podem dar-nos uma real dimensão do grau de contaminação, visto que as águas nem sempre apresentam um quadro definido, por causa das flutuações do fluxo [30 e 37].

Com efeito, a constante contaminação com metais pesados e não só, em áreas de mineração, é uma herança inevitável do desenvolvimento da sociedade humana, sobretudo desde a revolução industrial [38].

2.2 Métodos analíticos

Os métodos analíticos são uma importante ferramenta que nos permite conhecer a constituição da matéria e a quantidade de cada elemento que a constitui. Dependendo do tipo de material que vais ser alvo de análise, assim se deve escolher o método mais adequado e mais barato. Neste caso, foram utilizados a Espectroscopia de Emissão Óptica com Plasma Acoplado Indutivamente e, em alguns casos, a Colorimetria.

2.2.1 – Espectroscopia de Emissão Óptica com Plasma Acoplado Indutivamente (ICP-OES)

Na perspectiva de aumentar a precisão, baixar os limites de detecção e dosar simultaneamente vários elementos, algumas instituições de pesquisa e alguns laboratórios particulares optam pela Espectrofotometria de Emissão Óptica em plasma induzido (Inductively Coupled Plasma-Optical Emission Spectrometry – ICP-OES ou, simplesmente, ICP), para dosagem de elementos em extractos, ou na amostra total, de solos, plantas e os sedimentos de água. A utilização deste método intensificou-se a partir da década de 1970, devido, principalmente, aos avanços tecnológicos direccionados para as fontes de excitação e para as aplicações computacionais.

O princípio geral dos métodos espectrométricos baseia-se na excitação do átomo, que leva ao movimento de electrões de um orbital mais próximo, para outro mais afastado do núcleo, deixando o átomo num estado excitado. Quando a excitação é demasiado elevada, transforma-se o átomo excitado no respectivo catião, dizendo-se que ele atingiu o estado de ionização, o que a nível de análise dá problemas analíticos. Após o processo de excitação, os electrões dos átomos excitados e, ou, dos iões excitados, retornam rapidamente ao orbital de origem, emitindo energia electromagnética, fotões, com comprimento de onda específico para cada elemento e cada transição.

Segundo Raij [39], o plasma é a fonte de excitação mais efectiva para fins analíticos, sendo o árgon o mais utilizado. Os plasmas, por definição, são gases em que uma significativa fracção de átomos e moléculas se apresentam ionizadas [40]. As temperaturas do plasma atingem 10000 K, mas as análises são geralmente efectuadas na zona dos 6000 K. Estas temperaturas permitem que mesmo os electrões internos sejam excitados, o que é uma vantagem em relação aos outros métodos espectrométricos.

No espectrofotómetro, os fotões são transformados em sinais electrónicos, que são convertidos em concentração, após as devidas calibrações [41]. No ICP é permitido trabalhar com um conjunto de comprimentos de onda seleccionado para cada elemento. Dependendo do comprimento de onda escolhido, pode haver, ou não, interferências espectrais de outros elementos químicos, provocando distorções nos resultados analíticos. A interferência é detectada quando o pico de emissão (espectro) do elemento analisado não se apresenta uniforme.

A grandeza que normalmente é detectada pelos espectometros é a absorvância (A), equação (2.1). Todavia, é igualmente frequente fazerem-se registos de espectros com a unidade de transmitância, equação (2.2).

$$A = log\frac{I_0}{I_1} \tag{2.1}$$

$$T = \frac{I_1}{I_0} \tag{2.2}$$

Onde I_1 corresponde à intensidade de radiação transmitida e I_0 à intensidade de radiação incidente. Contudo, estas duas grandezas, absorvância e transmitância, podem ser relacionadas entre si através da equação (2.3).

$$A = \log(1/T) \tag{2.3}$$

Por outro lado, a equação (2.1) pode ser relacionada com o produto da concentração (c) pela espessura do meio absorvente, com a ajuda da constante Beer-Lambert (absorvitividade molar – ε), equação 2.4. Na qual a absorvância e a concentração estão relacionadas entre si, como se pode ver, também, pela equação 2.6, em que l – Espessura do meio absorvente, que no caso da EAA, será a chama.

$$\varepsilon lc = log\frac{I_0}{I_1} \tag{2.4}$$

Em que o ε será calculado a partir da equação 2.5, variando com a espécie absorvente (k) e com o comprimento de onda (λ).

$$\varepsilon = \frac{4\pi k}{\lambda} \tag{2.5}$$

$$A = \varepsilon lc \tag{2.6}$$

Em resumo, a lei explica que há uma relação exponencial entre a transmissão de luz através de uma substância, assim como também entre a transmissão e a longitude do corpo que a luz atravessa.

Se conhecemos I e ε, a concentração da substância pode ser deduzida a partir da quantidade de luz transmitida, equação 2.7.

$$c = log\frac{I_0}{I_1} \times \frac{1}{\varepsilon l} \qquad (2.7)$$

As unidade de c e ε dependem do modo como se expressa a concentração da substância absorvente. Se a substância é líquida, deve expressar-se como uma fracção molar.

Porém, a lei tende a não ser válida para concentrações muito elevadas, especialmente se o material dispersar a luz, sendo necessário proceder a diluições. [42 e 43].

2.2.1.1 – Problemas de Interferência

As interferências são um dos grandes problemas com que nos deparamos, tanto em técnicas tradicionais ou instrumentais, pois são uma das fontes de erro mais comum e, por vezes, são difíceis de identificar. No caso da espectroscopia de absorção atómica as interferências mais comuns podem ser espectrais ou não-espectrais.

A interferência espectral ocorre quando outras espécies existentes apresentam linhas espectrais próximas ou coincidentes com a do elemento a ser determinado, ou quando há dispersão da radiação dentro do espectrómetro. Enquanto as interferências não-espectrais remetem para problemas físicos da amostra.

2.2.1.1.1 Interferências espectrais

- Coincidência de linhas iónicas e atómicas – que acontece quando outras espécies têm linhas espectrais próximas ou coincidentes com o elemento a ser determinado, como se pode observar na figura 2.1.
- Dispersão da radiação dentro do espectrómetro - deve-se à reflexão da radiação nos componentes ópticos, originando imperfeições na grade de difracção.
- Produção de radiação de fundo – causada pelo próprio plasma, que é dependente do tipo de gás que forma o plasma, das espécies formadas e do processo cinético entre as espécies do plasma. Também pode ser causada pela energia de aquecimento excedente nas colisões e do arrefecimento das partículas.

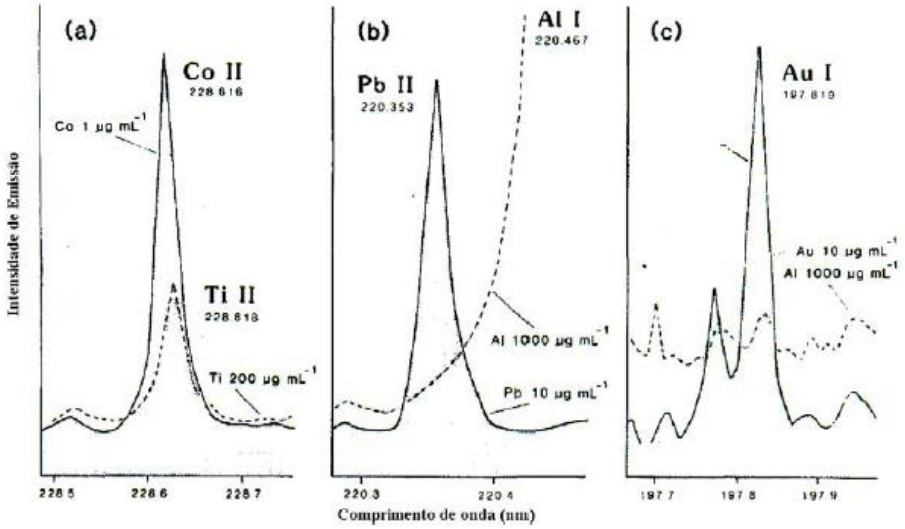


Figura 2.1 – Exemplos de interferências espectrais em ICP-OES. a) coincidência de linhas do Ti (228.618nm)e do Co (228.616 nm); b) sobreposição parcial da linha do A1 com a linha mais intensa do Pb e c) interferência da radiação de fundo (BG) na linha de emissão do Au, causado pela recombinação das linhas de emissão do Al. Adaptado de Jarvis e Jarvis [44].

2.2.1.1.2 Interferências não espectrais

As interferências não espectrais são todas as oriundas de mudanças de propriedades físicas da solução da amostra, que alteram a transferência de amostra para o plasma, temperatura, ou o numero de electrões no plasma [45] e podem estar ligadas com as seguintes situações:

- Alteração do sinal do analito por supressão;
- alteração das condições de nebulização e de excitação – deve-se, essencialmente, à variação do número de átomos resultante das mudanças de temperatura no ICP, causada pela variação da potencia RF e parâmetros físicos como o tipo e concentração do ácido na solução da amostra[46];
- perturbação na alimentação do nebulizador;
- perturbação na fase de dissociação molecular, tendo como consequência a presença de moléculas gasosas não dissociadas, que causam absorção molecular ou absorção não específica [47].

2.2.1.1.3 Formas de minimizar o efeito das interferências

A maioria das interferências encontradas nesta técnica pode ser reduzida ou completamente eliminada pelos seguintes procedimentos:

- Utilização de padrões e amostras de composição similar para eliminar os efeitos de matriz;

- Alteração da composição da chama ou a sua temperatura para reduzir a formatação de compostos estáveis na chama;
- Utilização de um método de correcção de linha de base[48];
- Utilização de espectrómetros com resolução superior, uma vez que estes permitem seleccionar linhas adicionais[45];
- Seleccionando uma linha alternativa que não sofra interferência[45, 49 e 50];
- Realizando uma correcção inter-elementar, em que a interferência é corrigida medindo-se a intensidade de emissão do elemento interferente noutro comprimento de onda;
- Criação de condições para realizar uma calibração multivariada, que se baseia no facto de que a radiação emitida pelo plasma é constituída pela adição de componentes individuais, para formar o espectro medido. Assim, os componentes são os espectros de emissão do elemento medido e o de fundo do plasma, pelo que se pode corrigir o espectro dos elementos ao subtrair o do fundo [45 e 51];
- Utilização de filtros ópticos colocados na fenda de entrada do monocromador ou detector, para evitar a dispersão da radiação dentro do espefoctómetro.

Nas interferências não-espectrais [45]:

- Correcção de matriz;
- Uso de padrão interno;
- Calibração por adição de padrão;
- Adição de surfactantes;
- Diluição da amostra;
- Separação de matriz;
- Correcção com modelos matemáticos;
- Evaporação do solvente presente no aerossol, mediante a introdução de volumes pequenos na câmara de nebulização, na ordem de μL[52].

2.2.2 – Colorimetria

A colorimetria, à semelhança da espectrofotometria atómica, pode definir-se como um procedimento analítico que permite determinar a concentração de espécies químicas através da absorção de energia radiante – luz. Assim, a colorimetria consiste numa técnica instrumental utilizada para a análise de amostras, que se baseia na comparação directa ou indirecta da intensidade da cor e

da qual se deduz a concentração, sendo a determinação realizada através de medições da sensação de cor [53].

Deste modo, é uma técnica recorrente na análise de quase todos os catiões, como zinco, e aniões, como fósforo, nitrito e nitrato, tendo o benefício de esta instrumentação básica ser relativamente pouco dispendiosa [54 e 55], obtendo-se resultados com uma boa sensibilidade e precisão. No caso das triazinas, o método colorimétrico pode ser tão ou mais eficiente do que o método cromatográfico de alta performance na detecção de resíduos [56]. O sinal do complexo colorido é convertido por fotometria, com ajuda de um fotomultiplicador, em voltagem, necessitando, por isso, o seu resultado de um grande controlo do pH, de um estado de oxidação específico, podendo, eventualmente, existir problemas com a interferência de outros materiais.

A principal vantagem dos métodos colorimétricos e espectrofotométricos é a de proporcionarem um meio simples para determinar quantidades diminutas de substâncias. O limite superior dos métodos colorimétricos, relativamente à gama máxima aconselhável para quantificação, corresponde à determinação dos constituintes em quantidades relativas inferiores a 1 ou 2% [57 - 59].

Esta técnica visa determinar a concentração de uma substância pela medida da absorção relativa de luz, tomando como referência a absorção da substância numa concentração conhecida. Pode-se utilizar uma fonte natural e/ou artificial de luz branca – fotometria do visível – ou uma célula fotoelétrica – fotometria do ultravioleta. No segundo caso, temos a eliminação, em grande parte, dos erros devidos às características pessoais de cada observador [54].

2.2.2.1 – Problemas de interferência

Interferentes ou interferências são todas as causas possíveis de originar erros sistemáticos, isto é, desvios em relação ao valor correcto, podendo ser apenas físicas ou químicas ou combinadas – físico-químicas. Contudo, os seus efeitos podem ser controlados através do uso de agentes inibidores que reagem com a substância interferente. Caso não se consiga obter o controlo desejado, pode ser necessário fazer uma separação preliminar do analisado da amostra complexa, ou destruir o complexo. Estão, ainda, disponíveis outras técnicas como a extracção, destilação, absorção em resina de troca de iões, precipitação e digestão.

2.2.3 – Limites analíticos

Com o melhoramento dos aparelhos que são utilizados para fazer as análises, a quantidade que é possível ser detectada também diminuiu. No entanto, alguns analitos são mais facilmente detectados por umas técnicas e outros por outras.

Contudo, o facto de serem detectados não significa que possam ser quantificados, por isso convém fazer a distinção entre limite de detecção (LD) e limite de quantificação (LQ). O primeiro corresponde à menor concentração do analito que pode ser detectável com confiança analítica (α=0.05, β=0.05)[4], mas não necessariamente quantificada. Por sua vez, o limite de quantificação (LQ) corresponde, nas condições experimentais estabelecidas, à menor concentração do analito que pode ser rigorosamente quantificada, com precisão e exactidão bem definidas (α≈0, β≈0) [59].

Para além destes, pode-se ainda definir o limite de decisão como a menor concentração de analito que apresenta uma probabilidade de 5 % de que o respectivo sinal possa ser considerado como sendo o sinal do branco e 50 % de hipóteses de se tratar, de facto, do analito (_α=0.05, β=0.5).

Tanto um, como o outro, podem ser estimados de três formas distintas: através de réplicas do branco, com base nos parâmetros da recta da curva de calibração e com base no desvio padrão do ajuste.

A estimativa com base em réplicas de branco ($m_0 \geq 10$) corresponde ao método recomendado pela IUPAC e ISO para avaliar os limiares analíticos inferiores.

As estimativas que se baseiam nos parâmetros da recta só devem ser utilizadas quando as anteriores conduzem a valores sem significado físico-químico ($X_{limiar} < 0$). Por outro lado, as que se baseiam no desvio padrão do ajuste, só devem ser utilizadas quando a estimativa anterior também conduz a valores sem significado físico-químico [60].

Tabela 2.1 - Resumo das estimativas dos limites da curva de calibração.

		Estimativas alternativas		
método de cálculo		1. Réplicas de brancos	2. Parâmetros da recta	3. Imprecisão do ajuste
limite de decisão	Y_D	$\overline{Y}_B + 1.65 \times \frac{s_B}{\sqrt{m_B}}$	$\theta_0 + 1.65 \times \sigma_{\theta_0}$	$\theta_0 + 1.65 \times \sigma_{\theta_0}$
limite de detecção	Y_{LD}	$\overline{Y}_B + 3.3 \times \frac{s_B}{\sqrt{m_B}}$	$\theta_0 + 3.3 \times \sigma_{\theta_0}$	$\theta_0 + 3.3 \times \sigma_{\theta_0}$
limite de quantificação	Y_{LQ}	$\overline{Y}_B + 10 \times \frac{s_B}{\sqrt{m_B}}$	$\theta_0 + 10 \times \sigma_{\theta_0}$	$\theta_0 + 10 \times \sigma_{\theta_0}$

No caso do ICP-OES, usado na análise das amostras referentes a este trabalho, os limites de detecção são os seguintes Fe-0.1 %; Ag-0.2 ppm; Ba, P, Cu, Cr, B, Zn, Ni, V, Mn, Co, Nb -10 ppm; Pb -

[4] As letras α e β referem-se às probabilidades unilaterais de o valor pertencer à distribuição do branco (valor de concentração nula) ou distribuição representativa do analito (concentração não nula), respectivamente.

20 ppm; Be – 1 ppm; Mo – 2 ppm; As – 20 ppm; Y – 5 ppm, Cd – 1 ppm, com uma precisão foi de 10%.

Para a colorimetria os limites de detecção são: 1ppm para Sn e Sb e 0.5 ppm para W e U e a precisão de 10%.

2.3 Tratamento de resultados

Numa fase inicial da quantificação é necessário explorar estatisticamente os dados referentes à curva de calibração do método utilizado, já que este procedimento permitirá converter os sinais instrumentais nas respectivas estimativas da concentração do analito em causa.

As normas ISO 8466-1 e 8466-2, referem-se concretamente ao ajuste polinomial da curva de calibração com modelos de primeiro grau [61] e segundo grau [62].

De acordo com estas normas, o método de estimativa dos parâmetros do modelo de calibração é realizado através da aproximação dos mínimos quadrados, por que os modelos polinomiais correspondem a funções linearmente dependentes sobre os parâmetros a estimar e, por isso, esta abordagem está estatisticamente correcta e devidamente fundamentada. Assim, para a calibração polinomial existem três questões essenciais que são determinantes para todo o tratamento posterior: a representatividade dos valores experimentais, a escolha do modelo e o diagnóstico de valores discrepantes.

Após efectuado o tratamento estatístico passou-se para uma análise mais pormenorizada dos dados, com a ajuda de métodos estatísticos mais robustos, como a Análise de Componentes Principais que nos permitem reduzir o número de variáveis em estudo e verificar quais são as mais importantes, ou a Análise por Mínimos Quadrados Parciais que possibilita observar se existe relações entre elas.

2.3.1 Representatividade dos valores experimentais

Uma vez que as estimativas paramétricas são sensíveis à variabilidade (incerteza) dos valores experimentais, estas devem reflectir, de uma forma o mais correcto possível, a precisão dos valores obtidos experimentalmente – os valores com maior significado estatístico (com menor incerteza associada) devem condicionar mais as estimativas paramétricas a obter.

Assim, antes de se começar a estimativa paramétrica é necessário determinar experimentalmente, dentro da gama analítica de trabalho, a homogeneidade das variâncias. Para tal,

recomenda-se a escolha de três níveis de concentração, ao longo da gama de trabalho: padrão de concentração mais baixo, padrão de maior concentração e padrão intermédio, onde se realizam dez (ou apenas cinco) réplicas desse padrão (n=10 ou n=5). Cada nível de concentração vai, deste modo, ser caracterizado por uma determinada incerteza (desvio padrão) que será posteriormente comparada através do teste F de Fisher.

Como hipótese inicial (hipótese nula, H0) assume-se que não há diferença estatística nas variâncias e como alternativa (hipótese secundária, H1) que estas estimativas são discordantes. O valor do teste (TV) calcula-se de acordo com a equação (2.8)

$$\mathrm{TV} = \frac{s_i^2}{s_j^2} \qquad (2.8)$$

Onde si e sj representam as variâncias de cada um dos padrões.

Sendo o teste estatístico bilateral, a estimativa de TV, equação (2.8), faz-se rearranjando o quociente das variâncias de modo a que este seja superior ou igual à unidade. Por este ser um passo determinante da curva de calibração, o nível de significância do teste corresponde a 0.01 (nível de confiança de 99 %) e o valor crítico tabelado pertence à distribuição de Fisher, com (n-1) graus de liberdade, no numerador e no denominador.

Se o valor TV não exceder este valor crítico, a hipótese nula (H0) é aceite, o que corresponde a dizer que existe homogeneidade da variância. Neste caso, a aproximação de ajuste por mínimos quadrados não ponderado (OLS) e válido. Caso contrário, os valores experimentais apresentam heterogeneidade da variância, o que implica que têm que ser utilizados os pesos estatísticos experimentais na estimativa (WLS).

Como a abordagem ponderada (WLS) é mais complexa, a norma ISO 8466-1 recomenda ainda que seja utilizado o padrão intermédio como termo de comparação, ou seja, faz o mesmo teste entre o primeiro padrão e o intermédio e entre o intermédio e o padrão mais elevado.

Agora, se a homogeneidade da variância for válida em ambos os casos, a calibração pode ser programada como correspondendo a duas calibrações lineares distintas – a calibração de gama baixa e a calibração de gama alta. A calibração de gama baixa envolve os padrões posicionados na gama de concentrações, situada entre o padrão mais baixo e o intermédio. Por outro lado, a calibração de gama alta abrange os padrões entre o valor intermédio e o padrão mais elevado. Caso a homogeneidade da variância continue a tender para o caso heterocedástico, sobretudo na gama de concentração mais elevada, deve-se reduzir a gama analítica a gama mais baixa ou então, utilizar o método de estimativa ponderada.

2.3.2 Escolha do modelo

Importa agora saber qual o modelo que melhor se coaduna aos valores experimentais obtidos.

As normas ISO 8466-1 e 8466-2, apenas assumem a discussão estatística entre dois tipos de modelos polinomiais de primeiro grau, equação (2.10) e de segundo grau, equação (2.11),

$$P_1: \hat{Y}_1(i) = a_{01} + a_{11}.X_i \quad (2.9)$$

$$P_2: \hat{Y}_2(i) = a_{012} + a_{12}.X_i + a_{22}.X_i^2 \quad (2.10)$$

Onde Ŷ representa o valor da variável dependente, estimado pelo modelo polinomial de primeiro grau (1) ou de segundo grau (2).

Segundo as normas ISO 8466-1 e 8466-2, deve-se comparar o desempenho dos modelos no ajuste dos valores experimentais (Yi), através das respectivas somas de quadrado (SS). A variância do ajuste e dada pelo quociente da equação (2.11):

$$\sigma_{fit}^2 = \frac{SS}{\nu} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{(n-p)} \quad (2.11)$$

Onde v representa o número de graus de liberdade, n e p são o número de pontos ajustados e p o número de parâmetros do modelo (p = 2 no caso P1 e p = 3 no caso P2).

No teste de Mandel, para a escolha do modelo, assume-se como hipótese inicial (H0) que ambas as funções (P1 e P2) ajustam bem os valores experimentais e, como hipótese alternativa (H1), que o polinómio de segundo grau representa melhor os valores experimentais. Neste caso o valor do teste a calcular é dado pela expressão da equação (2.12),

$$TV = \frac{\Delta\sigma^2}{\sigma_{pe}} = \frac{(SS_1 - SS_2)}{\sigma_{fit}^2} \quad (2.12)$$

Onde SS representa a soma de quadrados obtida com o ajuste polinomial de primeiro (1) e segundo (2) graus. O valor obtido (TV) deve ser comparado com o valor crítico da distribuição unilateral de Fisher ao nível de confiança de 99 %, referente ao número de graus de liberdade de (1) e (n – p – 3), respectivamente. Se o valor obtido não exceder este valor critico $F_{0.01}^{u}(n-1, n(m-1))$.

2.3.3 Teste de valores discrepantes

2.3.3.1 Teste de Grubbs

No teste de Grubbs a equação (2.13) objectiva a comparação entre o desvio do valor suspeito e a média de todos os valores, com o desvio padrão das amostras.

$$G = \frac{\left| x_? - \bar{x} \right|}{s_x} \tag{2.13}$$

De seguida compara-se este valor com um valor crítico tabelado, para um intervalo de confiança de 95%, sendo apresentada a tabela A1.1 do Anexo A1. Caso o G seja superior ao valor crítico, então estamos na presença de um outlier e é necessário eliminar esse valor, de modo a que este não influencie incorrectamente os resultados e tratamentos estatísticos.

2.3.3.2 T-student

Desenvolvido por William Sealy Gosset, inicialmente para o estudo do comportamento de um pequeno número de amostras, a distribuição *t* de Student, é uma distribuição de probabilidade estatística e teórica, sendo simétrica, campaniforme e semelhante à curva normal padrão. Todavia, possui caudas mais largas, ou seja, uma simulação da t de Student pode gerar valores mais extremos que uma simulação da normal, e é normalmente usado quando está em causa a comparação de estimativas de posição e assumindo que se trata de distribuições normais e independentes.

Na verdade, outra forma de verificar a existência de outliers consiste em confirmar se um determinado valor duvidoso, ($x_?$), pertence a um grupo de valores representados pela media destes, $\bar{x}$, através deste teste, equação (2.14),

$$TV = (t) = \frac{\left| x_? - \bar{x} \right|}{s_x} \tag{2.14}$$

Onde s_x corresponde à dispersão calculada após a remoção do valor questionado do conjunto de valores. A probabilidade do valor duvidoso não ser um outlier pode ser representada através duma ferramenta do Excel denominada “tdist”, onde o valor de prova é comparado com o valor crítico, representado na tabela A1.2 do Anexo A1, a um nível de confiança de 99 %. Caso o valor de prova seja inferior ao valor crítico é um outlier.

2.3.3.4 Teste de Fisher

Por vezes, é necessário comparar, estatisticamente, os desvios padrão de diferentes resultados, recorrendo-se, então, ao teste-F. Para percebermos a utilização deste método importa explicar aqui duas situações possíveis, o teste unilateral e o teste bilateral. Quando estamos a comparar dois métodos, em que não existe qualquer expectativa sobre o desvio padrão de um em relação ao outro, ou seja, se um é significativamente maior ou menor que o outro, estamos perante um teste bilateral. No caso de, por exemplo, querermos saber se um método novo é mais preciso que um método de referência, em que esperamos que o desvio padrão do método de referencia seja significativamente maior que o do método novo, então estamos perante um teste bilateral.

O teste-F obtém-se pela equação (2.15) e esse valor é posteriormente comparado com as Tabela A1.3 e A1.4, no Anexo A1, onde para a hipótese nula ser válida, o valor de F tem de ser menor que o valor de F crítico, de acordo com os respectivos graus de liberdade e intervalo de confiança.

$$TV = (F) = \frac{s^2{}_x}{s_x} \tag{2.15}$$

2.4 Análise de Componentes Principais: PCA

Inventado em 1901, por Karl Pearson [63], a Análise de Componentes Principais ou PCA, do inglês "Principal Component Analysis", tem, hoje em dia, por finalidade básica a redução de dados a partir de combinações lineares das variáveis originais [16], isto é, tem a função de ferramenta para a análise de dados exploratórios e para a criação de modelos de previsão.

Normalmente, os resultados obtidos através do PCA são discutidos em função dos componentes dos dados auferidos e respectivo peso. Neste sentido, esta técnica é a mais simples no que toca à análise, baseada em dados multivariados, visto que, muitas vezes, a sua operação pode ser pensada como uma revelação da estrutura interna dos dados de uma forma que melhor os expõe.

2.4.1 Definição de PCA

É matematicamente definido como uma transformação linear ortogonal, ou seja, altera os dados para um novo sistema coordenado, para que a primeira maior variância de qualquer projecção de dados se aloje na primeira coordenada (ou primeira componente principal), enquanto a segunda

maior variância se aloja na segunda coordenada e assim sucessivamente, de forma que as novas variáveis são funções lineares das variáveis originais. Com efeito, a primeira componente principal considera a maior variabilidade possível nos dados, e cada componente sucedente considerará, desta forma, toda a variabilidade restante possível.

Assim, dependendo do tipo de aplicação, pode também ser apelidado de Transformador de Karhunen-Loève (KLT) [64], Transformador de Hotelling ou Decomposição Ortogonal Própria (POD). Deste modo, o PCA é, teoricamente, o transformador, por excelência, de dados fornecidos nos menores termos de uma recta/raiz quadrada.

Na verdade, o PCA estima a estrutura de correlação das variáveis, cuja importância num modelo de PC é indicada pelo tamanho da sua variância residual, comummente utilizado na selecção de variáveis. Na continuação desta linha de raciocínio, a figura 2.2 mostra esquematicamente a separação entre os dados estruturados e o resíduo.

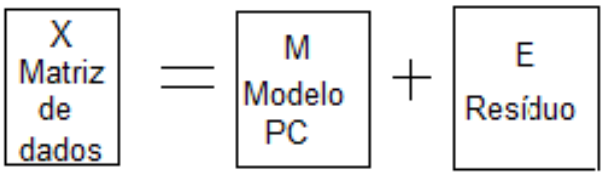


Figura 2.2: A matriz X obtém-se a partir da combinação da estrutura subjacente do modelo PC (M) e do resíduo/ruído (E). Por sua vez, a estrutura subjacente resulta da antecipação ou estimativa, a partir da matriz X.

A figura 2.3 demonstra as propriedades da projecção PCA, sendo perceptível que, com uma interpretação adequada das projecções, pode verificar-se quais as características dominantes dadas pelo conjunto multivariado do grupo de dados.

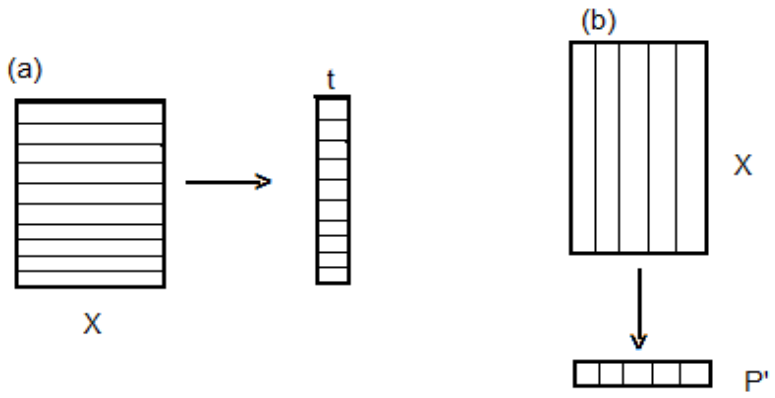


Figura 2.3: a) a projecção da matriz X no vector t; b) Projecção da matriz X no vector P'.

A projecção da matriz X no vector t é o mesmo que assumir um número para todos os objectos, isto é, para cada uma das linhas. A projecção é escolhida de forma que os valores em t tenham propriedades desejadas e que o ruído contribua o mínimo possível.

Ao fazer-se a projecção da matriz X no vector p' atribui-se escalamento a todas as variáveis de cada coluna. A projecção é escolhida de forma que os valores em p' tenham as propriedades desejadas e o que erro tenha uma contribuição mínima.

Para melhor se entender as informações anteriores, pode-se recorrer à ilustração da figura 2.4, onde se observa uma matriz (3x4) [65] a passar por várias etapas ao longo do seu estudo.

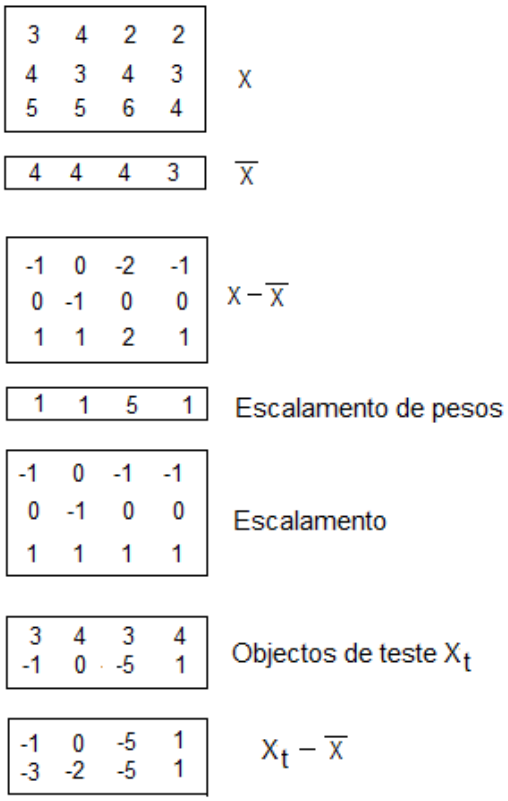


Figura 2.4: Matriz de dados usada como exemplo. Inclusão de duas operações extra ao conjunto de teste, a Média – centragem e a variância – escalamento.

2.4.2 Vantagens

Teoricamente, o PCA é um excelente esquema linear, em termos de menor erro de raiz quadrada da média, permitindo a transformação de um conjunto de vectores dimensionalmente altos para um conjunto de vectores dimensionalmente baixos, podendo reconstruir o conjunto original, mais tarde. É, basicamente, uma análise não paramétrica, sendo uma resposta única e independente,

que não tem qualquer hipótese sobre a probabilidade de distribuição dos dados.

De facto, o PCA permite a simplificação, redução de dados, modelação, detecção de outliers, selecção de variáveis, classificação, predição e separação. Contudo, nem sempre é necessário recorrer a todas estas operações, uma vez que, consoante o problema inicial, assim se escolherá os diferentes cálculos a ser aplicados à matriz de dados, de forma a se poder realizar a melhor análise.

Como tal, pode dizer-se que engloba o cálculo da decomposição de valores próprios da matriz covariante dos dados ou a decomposição de valores singulares de dados numa matriz, especialmente depois de centrar a média dos dados para cada valor.

Na verdade, o ponto de partida para a análise de dados multivariados é dado pela matriz de dados, isto é, a tabela de dados, que se encontra identificada por um X na figura 2.5. As linhas N da tabela são designadas como "objectos", correspondendo às amostras, enquanto as colunas K se denominam como "variáveis", compreendendo as medidas feitas aos objectos, figura 2.5.

Portanto, esta figura permite dar uma ideia das matrizes e vectores utilizados no PCA, visto que muitos dos cálculos aqui usados têm como finalidade encontrar relações entre objectos, podendo tornar-se bastante interessantes, uma vez que se existe a possibilidade de formar classes entre objectos similares. Para além disso, apesar da classe dos elementos já ser conhecida, pode, também, ser feita uma exploração dos dados avaliados, de modo a poder associar-se a detecção de outliers a este processo, já que não existem valores que por vezes não pertencem a nenhuma classe conhecida [66 e 67].

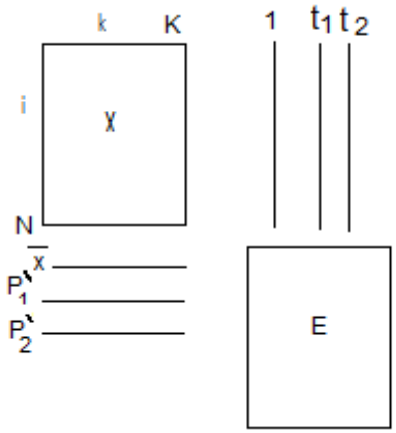


Figura 2.5: Matriz de dados X com as duas componentes principais. O índice i é usado para os objectos (linhas) e o índice k para variáveis (colunas). Tem N objectos e K variáveis. A matriz E contém os resíduos, que equivalem aos dados que não são explicados pelo modelo PC.

O esquema da figura 2.5 pode ser matematicamente representado pela equação 2.16

$$X = 1\bar{x} + TP' + E \qquad (2.16)$$

Como já foi referido, outro cálculo disponível é o da redução dos dados, que se torna muito útil quando há grandes quantidades de dados que podem ser aproximadas por uma estrutura modelo pouco complexa.

No geral, a maioria das matrizes de dados pode ser simplificada com esta técnica, até porque uma grande tabela de números é uma das coisas mais difíceis para a compreensão da mente humana. Todavia, pode aliar-se as capacidades do PCA a um conjunto muito bem seleccionado de objectos e variáveis, de forma a construir um modelo de comportamento de um sistema físico ou químico que vai ter como finalidade a previsão de novos dados que são medidos para o mesmo sistema. Acresce ainda o intuito de esta técnica poder ser utilizada na análise de sistemas com mobilidade, como a dissolução constante de algumas misturas, apelidada de curva de resolução [66 e 67].

2.4.3 Desvantagens

As últimas duas propriedades, construção de modelos comportamentais e o estudo de curvas de dissolução, são consideradas fortes e fracas ao mesmo tempo, uma vez que, por ser uma técnica "não paramétrica", não pode incorporar nenhum tipo de conhecimento superior. Deste modo, as compressões de PCA, estão usualmente sujeitas a perdas de informação.

Para além disto, a aplicabilidade do PCA é limitada a suposições criadas na sua derivação e às características que lhe são inerentes, que são as seguintes [68]:

- **Suposição na linearidade**

  Assume-se, à partida, que o conjunto de dados observados é uma combinação linear de certezas básicas, de forma que os métodos não lineares, como o núcleo do PCA, têm vindo a ser desenvolvidos sem se assumir a variabilidade.

- **Suposição da importância estatística da média e covariância**

  O PCA usa vectores próprios da matriz de covariância, encontrando, apenas, eixos independentes dos dados, conforme a suposição Gaussiana. Para uma associação não Gaussiana ou para dados Gaussianos multi-modais, o PCA limita-se a correlacionar os eixos. Quando o PCA é usado para agrupar, deparamo-nos com a sua maior limitação: a incapacidade de não separar classes, uma vez que não dá uso a tabelas de classificação de pontos característicos dos vectores. Não há nenhuma garantia, portanto, que as direcções de variância máxima contenham bons pontos característicos para a discriminação.

- **Suposição de que grandes variâncias têm dinâmicas importantes**

  O PCA desempenha, tão-somente, uma rotação coordenada que alinha os eixos, transformados com as direcções da variância máxima. Isto é, quando se verifica que os dados observados têm uma grande razão sinal – ruído, em que as componentes principais com grandes variâncias correspondem a dinâmicas de interesse e as com variâncias mais baixas ao ruído.

Essencialmente, o PCA envolve escalamento e rotação, sendo que as suposições acima, foram criadas para o cálculo algébrico do conjunto de dados, que têm vindo a ser desenvolvidos noutros métodos sem nenhuma destas suposições.

2.5 Mínimos Quadrados Parciais: PLS

Introduzido por Wright em 1920 [64], o caminho de análise e modelagem causal foi, mais tarde, desenvolvido, no final da década de 1960, por Herman O. A. Wold [66]. Inicialmente, a regressão parcial de mínimos quadrados era mais utilizada no campo da econometria, vindo a ser adoptado pela área de química para uso em química analítica, física e estudos clínicos, anos depois [67]. Ao criar os mínimos quadrados parciais, Wold tinha como objectivo a abordagem da teoria do elo fraco e dos dados fracos [65, 66 e 68].

Assim, os **Mínimos Quadrados Parciais**, ou **PLS**, do inglês Partial Least Squares, tem por finalidade a previsão ou análise de um conjunto de variáveis dependentes, a partir de um grupo de variáveis independentes ou predictores, do qual se extrai um conjunto de factores ortogonais, apelidados de variáveis latentes, que tem o melhor poder de previsão.

A análise Parcial de Mínimos Quadrados é uma técnica estatística multivariada que permite a comparação entre as variáveis de resposta múltipla e múltiplas variáveis explicativas. Deste modo, os mínimos quadrados parciais são um de uma série de covariâncias baseadas em métodos estatísticos que são muitas vezes referidas como a modelagem de equações estruturais ou SEM, que foi projectado para lidar com a regressão múltipla quando tem uma pequena amostra de dados, os valores em falta, ou multicolinearidade. A regressão parcial de mínimos quadrados foi demonstrada em vários casos com dados reais e em simulações [69 e 70]. Na altura, esta foi uma técnica muito popular na ciência dura, especialmente na química e na quimiometria, onde há um grande problema com um elevado número de variáveis correlacionadas e um número limitado de observações, e apesar de também ser usada em marketing, a sua utilidade real é mais limitada, embora os dados aí

existentes tenham problemas semelhantes [69].

2.5.1 Definição de Mínimos Quadrados Parciais

O termo, mínimos quadrados parciais significa, especificamente, o cálculo dos mínimos quadrados óptimos que se integram em parte numa matriz de correlação ou matriz de covariância [65 e 71]. A técnica de Mínimos Quadrados Parciais destina-se assim, a lidar com os problemas nos dados, mais especificamente, no que diz respeito a conjuntos de dados pequenos, a valores em falta e a multicolinearidade. Em contrapartida, mínimos quadrados ordinários (OLS – ordinary least squares) produzem resultados de regressão instáveis quando os dados provêm de um tamanho reduzido de amostra, se houver valores em falta e multicolinearidade entre predictor na regressão OLS, aumenta o erro padrão dos coeficientes estimados [72]. A multicolinearidade elevada aumenta o risco de, teoricamente, o predictor (inteiro) sem ser rejeitado, a partir do modelo de regressão como variável não significativa.

O objectivo de mínimos quadrados parciais é prever X e Y para descrever a estrutura subjacente comum às duas variáveis [73]. Os mínimos quadrados parciais são, deste modo, um método de regressão que permite a identificação dos factores subjacentes, que são uma combinação linear das variáveis explicativas ou X (também conhecido como variáveis latentes) que será melhor modelo da resposta ou variáveis Y [74]**.**

Apesar de ser similar à análise de regressão de componentes principais, à análise canónica e aos mínimos quadrados alterados, considera-se que seja a melhor alternativa à regressão linear e a métodos de regressão de PCA, uma vez que prevê mais parâmetros de modelos robustos que não muda com novas amostras de calibração da população [64 e 66]. Acresce ainda dizer que PLS é uma actualização melhorada do PCA, uma vez que, a solução deriva dos Mínimos Quadrados Parciais e é limitada à parte da matriz covariante que está directamente relacionado com a manipulação experimental ou com o comportamento [75].

A parte da correlação ou da matriz de covariância permite verificar se os mínimos quadrados estão aptos para a correlação entre as variáveis exógenas ou “X” e entre as medidas ou variáveis dependentes “Y”. Pelo menos, de forma parcial, entre dois ou mais blocos de variáveis e criando um novo conjunto de variáveis que é optimizado para covariância máxima (correlação não máxima), utilizando o menor número de dimensões [73].

2.5.2 Vantagens

- Capacidade de modelar múltiplas variáveis, tanto dependentes como independentes.
- Pode tratar multicolinearidade em IVs.
- Despiste robusto do ruído dos dados e da ausência dos mesmos.
- Cria latentes independentes directamente sobre a base de produtos cruzados envolvendo resposta(s) variáveis = previsões fortes.
- Permite latentes reflexivas e formativas.
- Aplicável em amostras pequenas.
- Distribuição livre.
- Faixa de punho de variáveis: nominal, ordinal, contínua [73].

2.5.3 Desvantagens

- Dificuldade em interpretar os carregamentos das variáveis latentes independentes (baseadas em relações CrossProduc com respostas não variáveis, como na análise factorial convencional, a correlação entre os manifestos independentes).
- Não reconhece propriedades distributivas das estimativas:
  - Não consegue ter significado, a não ser de inicialização.
- Falta de estatísticas de teste modelo [73].

Capítulo 3

3.1 Procedimentos experimentais

Por definição, a amostragem é o processo assente na recolha representativa de um todo, de acordo com um plano definido pelo tipo e qualidade do determinado material ou amostrado, existindo várias técnicas que podem ser aplicadas, como a amostragem aleatória simples, sendo a escolha desta determinada pelo propósito e condições da amostragem. Deste modo, procura-se que a amostra seja uma fracção representativa de um todo, seleccionado de tal modo que possua as características essenciais do conjunto que ela representa. Esta técnica, e, em particular, os seus processos, aplicam-se em diversas áreas do conhecimento, constituindo, muitas vezes, a única forma de obter informações sobre uma determinada realidade que importa conhecer.

Já a sua teoria estuda as relações existentes entre uma população e as amostras dela extraídas, sendo muito útil para a avaliação de grandezas desconhecidas da população, ou para determinar se as diferenças observadas entre duas amostras são devidas ao acaso ou antes verdadeiramente significativas.

3.1.1 Processo de amostragem

Por vezes, o processo de amostragem nem sempre é uniforme, porque a forma de realizar o trabalho de amostragem pode variar consoante a finalidade a que este se destina [76 e 77].

3.1.1.1 Amostra em processo

O processo deverá ser realizado por pessoal devidamente preparado nos aspectos operacionais e de segurança. De seguida, devem-se seguir as etapas previamente estabelecidas, utilizando acessórios e recipientes já definidos, descontaminado e isentos de interferências, para a colecta das amostras, com o produto em quantidade suficiente para realização de todos os ensaios necessários. A amostra do produto deverá ser devidamente rotulada para garantir a identificação e a restreabilidade do mesmo. As amostras deverão ser, então, disponibilizadas para análise e retenção, conforme procedimento interno da instituição.

3.1.1.2 Amostragem de Produto Acabado

Este processo deverá ser, igualmente, realizado por pessoal devidamente qualificado. A recolha do produto acabado tem de ser efectuada após o envase, em quantidade e periodicidade suficientes

para atender às necessidades de controlo.

3.1.1.3 Processo utilizado para a recolha e tratamento das amostras

Para a recolha com vista à prospecção geoquímica, - que foi o objectivo com que o ex-IGM fez esta amostragem -, cartografia geoquímica e determinação de fundos regionais, há consenso para a recolha e tratamento das amostras de sedimentos de correntes. As amostras são colhidas antes de cada confluência de linhas de água e o número colhido depende da densidade de amostragem pretendida. No presente caso a área tinha 171 km^2 e foram colhidas seiscentas e noventa e três amostras. Por desconhecer o procedimento que foi empregue no tratamento das amostras, não me é possível descreve-lo, contudo, o procedimento utilizado no Laboratório Químico, do Departamento de Ciências da Terra, da Universidade de Coimbra é o descrito em baixo [78].

As amostras de sedimentos são colhidas antes da confluência das linhas de água, procedendo-se à colheita do material, que tem entre 1 Kg e 1.5 Kg, para um saco de plástico transparente, devidamente identificado com o nome da amostra e a data da amostragem. Findo este processo, transportam-se as amostras para o laboratório, onde são secas na estufa a uma temperatura que deve ser inferior a 40 ^{0}C. É-lhes, então, feita uma limpeza de folhas, ramos e de outro material mais grosseiro, para poder ser peneirada, de forma mais eficiente, com um crivo (peneiro) de 2 mm. A fracção superior é eliminada, enquanto a outra passa para a fase seguinte, onde sofre homogeneização e quartiamento, ficando dividida em metades. Aqui, uma é preservada num saco identificado, enquanto a outra é, novamente peneirada, mas com um crivo de 250 µm. A fracção superior é guardada e identificada, já a inferior é, novamente, homogeneizada, quartiada, removendo-se uma porção de 100 g, da qual se retira uma pequena parte para a realização da digestão ácida.

Para execução da digestão ácida pega-se na fracção seleccionada, pesa-se 0.5 g, deitando-se para um tubo de digestão, onde se adiciona ácido nítrico e ácido clorídrico, com um grau de pureza de 65 % e 37 %, respectivamente, numa porção de 1:3, por esta ordem, evitando que a reacção ocorra de forma violenta. A solução repousa durante a noite, para no dia seguinte se aumentar a temperatura do bloco de digestão até atingir os 140 ^{0} C, de forma faseada, impendido, assim, que aconteça uma subida exponencial da temperatura, o que causaria uma digestão menos eficaz. Uma vez atingidos os 140 ^{0}C, o bloco permanece a essa temperatura durante duas horas, ao fim das quais se desliga o bloco, deixando-se arrefecer até à temperatura ambiente.

O próximo passo consiste em filtrar a solução resultante da digestão, com um filtro de 0.45 µm, evitando, assim, que as partículas que não foram digeridas entrem nos aparelhos, utilizados para

a análise das amostras, entupindo as tubagens e causando interferências nos espectros. Deste modo, deita-se o "licor" obtido no processo para um balão de diluição de 100 ml, com a ajuda de um funil, com um filtro incorporado. Lava-se o tubo com uma solução de ácido nítrico a 0.5 M, que depois se junta ao produto da digestão, novamente através do filtro, aferindo-se o balão.

Antes destas amostras serem analisadas têm de ser diluídas com uma percentagem de água, para que não ocorra a saturação espectral do aparelho e o efeito do ácido nas tubagens seja reduzido.

3.1.1.3.1 Equipamentos e reagentes normalmente utilizados neste processo

Equipamentos:

- Balança de precisão;
- Bloco de Aquecimento;
- ICP-OES, da marca Horiba Jobin Jvon, modelo JY 2000-2.

Reagentes:

- Ácido Clorídrico a 37 %, da marca Riedel-De Haën;
- Ácido Nítrico a 65 %, da marca Merck;
- Padrões Multi-elementares (As, Nb, Sn,Ta, Ti, W) e (Th, U, V), da INORGANIC Ventrures ;
- Padrão Multi-elementar (Ag, Al, B, Ba, Bi, Ca, Cd, Co,Cr, Cu, Fe, Ga, In, K, Li, Mg, Mn, Na, Ni, Pb, Sr, Ti, Zn), da marca Merck.

3.2 Tratamento de resultados

3.2.1 Preparação da curva de calibração

A curva de calibração é um método geral para determinar a concentração de uma substância numa amostra desconhecida, através da sua comparação com um conjunto de amostras padrão de concentração conhecida. Ou seja, a curva de calibração representa a forma como a resposta instrumental "o sinal analítico" se adapta a mudanças com concentração desse analito (substância a ser medida).

Para se obter a curva de calibração é necessário preparar um conjunto de padrões com concentrações crescentes, do o(s) analito(s) que vai(ão) ser analisado(s), de forma a que as concentrações mais baixa e mais alta fiquem dentro da gama de trabalho, previamente estabelecida. De tal forma que se possa obter uma equação matemática para explicar esta relação e, a partir daí, encontrar o(s) valor(es) do analito(s) nas amostras desconhecidas. Sendo que esta equação matemática não é mais do que a equação da recta que melhor aproxima os pontos resultantes da representação gráfica do sinal, em função da concentração, equação (3.1) e onde (y) representa a intensidades de sinal, x a concentração, o m a inclinação da curva e b a ordenada na origem.

$$y = mx-b \quad (3.1)$$

Uma vez obtidos o(s) sinal(is) (y) correspondente(s) ao(s) analito(s) basta substituir esse valor na equação 3.1 e rearranja-la de forma a obter-se o valor da concentração (c) [79].

Contudo, nem sempre se consegue uma relação linear entre todos sinais obtidos dos aparelhos e a respectiva concentração de analito, já que a resposta instrumental é em geral, altamente dependente da condição do analito, dos solventes usados e de factores externos. Como tal, alguns pontos da recta de calibração fogem a essa linearidade, sendo, por vezes, necessário remover esses pontos, a fim de se obter uma melhor relação. Logo, é conveniente utilizar na preparação da recta de calibração, pelo menos cinco padrões de concentrações espaçadas entre si.

3.2.1.1 Réplicas de brancos

Para que se possa saber se um método é ou não muito sensível a um determinado analito, deve-se encontrar os limiares analíticos, ou seja, o limite de detecção e o limite de quantificação.

Existem três formas de determinar estes limites, como descrito na norma ISO 8466-1, contudo, o mais utilizado quando se pretende fazer a validação de um método, consiste em utilizar uma série de amostras de branco, pelo menos dez, preparadas independentemente, (ou um padrão de baixa concentração, caso o branco não tenha flutuação significativa). Depois de se obter o sinal do analito em cada amostra de branco, utiliza-se a equação da recta de calibração para se obter a respectiva concentração. O passo seguinte consiste em calcular a média (x_0) e desvio padrão (s_0) das concentrações do analito nos brancos se substituir nas equações (3.2) e (3.3).

$$LD = x_0 + 3{,}3s_0 \quad (3.2)$$

$$LQ = x_0 + 10s_0 \quad (3.3)$$

No entanto, de cada vez que é feita uma nova curva de calibração, os métodos mais recentes e, em especial o ICP-OES existente aqui no laboratório, permite o cálculo do limite de detecção e o limite de quantificação, a partir da recta de calibração, isto é, a partir da estatística de mínimos quadrados da recta de calibração. Neste caso, admite-se que o desvio-padrão da estimativa (Sy/x) representa o desvio-padrão do branco, como nas equações (3.4) e (3.5).

$$LD = \frac{3.3 s_0}{declive} \quad (3.4)$$

$$LQ = \frac{10 s_0}{declive} \quad (3.5)$$

O LQ estimado deverá ser confirmado experimentalmente com um padrão de concentração semelhante.

3.2.1.2 Réplicas de padrões

Os métodos analíticos são essencialmente comparativos, já que a grandeza medida, em laboratório, deve ser relacionada com os respectivos valores do padrão utilizado na calibração.

Numa fase inicial do trabalho analítico, a resposta do equipamento quantificador (yi) é avaliada com base na submissão de um certo conjunto de padrões de concentração conhecida (xi) que garantem a rastreabilidade do processo.

Todavia, para garantir que os valores obtidos para os padrões que compõem a curva de calibração estão correctos, pode fazer-se várias réplicas de cada um dos padrões (no mínimo 3) e o valor final será dado pela média e respectivo desvio padrão.

Apesar de a maioria das curvas de calibração pertencerem a regressões lineares (y = mx+b), existem casos em que o melhor modelo para a calibração parece não ser linear, mas sim pertencer à regressão não linear, onde a curva é dada por uma equação de segundo grau, $y = ax^2+bx+c$, ou seja, há elementos com mais peso do que outros.

Para se saber qual dos modelos é o mais correcto recorre-se aos estimadores paramétricos mínimos quadrados não ponderados (OLS) e mínimos quadrados ponderados (WLS).

a) Mínimos quadrados não ponderados (OLS)

Neste estimador, todos os elementos da amostra estão igualmente representados na estimativa, ou seja, o equipamento apresenta uma resposta linear dentro da gama de

concentrações dos padrões (X_1 a X_N). O sinal obtido y_i pode ser descrito através da equação (3.1) e a variância do ajuste, pela equação (3.6),

$$\sigma^2_{fit} = \frac{SS}{(n-p)} = \frac{\sum_{i=1}^{n}(y_i - b_0 - m.x_i)^2}{(n-p)} \qquad (3.6)$$

sendo as respectivas equações normais,

$$\begin{cases} \frac{\partial SS}{\partial b_0} = 2\times(\sum_{i=1}^{n}(y_i - b_0 - m.x_i))\times(-1) = 0 \\ \frac{\partial SS}{\partial m} = 2\times(\sum_{i=1}^{n}(y_i - b_0 - m.x_i))\times(-x_i) = 0 \end{cases} \qquad (3.7)$$

Rearranjando, obtém-se o seguinte sistema de equações (3.8), representado na forma de matriz ampliada

$$\left\{A \mid ind\right\} = \left\{\begin{array}{cc|c} N & S_x & S_{x^2} \\ S_x & S_{x^2} & S_{xy} \end{array}\right\} \qquad (3.8)$$

sendo $S_x = \sum_{i=1}^{n} x_i$, $S_y = \sum_{i=1}^{n} y_i$, $S_{x^2} = \sum_{i=1}^{n} x^2{}_i$, $S_{xy} = \sum_{i=1}^{n} x_i \times y_i$. Contudo, este sistema de equações só possui solução se a matriz de projecto possuir determinante, ou seja, ($\Delta = N{:}S_{x2} - S_x.S_x \neq 0$).

O declive (m) pode ser estimado através da regra de Cramer, equação (3.9)

$$m = \frac{\Delta_1}{\Delta} = \frac{N \times S_{xy} - S_x \times S_y}{N \times S_{x^2} - S_x \times S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y}).}{\sum_{i=1}^{n}(x_i - \bar{x})^2.} \qquad (3.9)$$

e a intercepção com o eixo das ordenadas (b) através da equação do modelo que satisfaz o ponto "centróide" ($\bar{x}$; $\bar{y}$) da curva da equação (3.10),

$$b = \bar{y} - m\bar{x} \qquad (3.10)$$

Invertendo a matriz de projecto (A) obtém-se a matriz de covariância (C) que permite estimar as dispersões das estimativas paramétricas, enquanto que a imprecisão paramétrica (OLS) pode ser

estimada através da equações (3.11) e (3.12),

$$\sigma_b^2 = \sigma_{fit} \times \sqrt{c_{11}} = \sigma_{fit} \times \sqrt{\frac{S_{x^2}}{\Delta}} = \sigma_{fit} \times \sqrt{\frac{\sum_{i=1}^{n} x^2}{N \times \sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad (3.11)$$

$$\sigma_b^2 = \sigma_{fit} \times \sqrt{c_{22}} = \sigma_{fit} \times \sqrt{\frac{N}{\Delta}} = \sigma_{fit} \times \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad (3.12)$$

Em que $\Delta = (N \times \mathrm{Sx}^2 - \mathrm{Sx} \times \mathrm{Sx}) = N \times \sum_{i=1}^{n} (x_i - \bar{x})^2$. e N corresponde ao número de padrões da curva de calibração que foram utilizados no ajuste dos parâmetros da recta de calibração, devendo incluir o branco.

A análise das amostras deve ser sempre efectuada com base em réplicas (M ensaios independentes para cada amostra i).

Antes de se efectuar a estimativa da concentração na amostra deve-se começar por verificar se são válidos os pressupostos para a estimativa paramétrica da média – ausência de erro sistemático e de valores discrepantes, erro aleatório independente e com distribuição normal.

Após esta verificação, as estimativas a efectuar estão nas equações (3.13) e (3.14),

$$\sigma_{\bar{y}} = \frac{\sigma_{y_s}}{\sqrt{M}} = \frac{\sum_{i=1}^{M} y_{s_i}}{\sum_{i=1}^{M}} = \frac{S_{y_s}}{M} \quad (3.13)$$

$$\sigma_{\bar{y}} = \frac{\sigma_{y_s}}{\sqrt{M}} = \sqrt{\frac{\sum_{i=1}^{M} (y_{s_i} - \overline{y_s})^2}{M(M-1)}} \quad (3.14)$$

sendo posteriormente a concentração estimada através dos parâmetros da recta de calibração, equação (3.15)

$$\bar{x}_s = \frac{\overline{y_s} - b}{m} \quad (3.15)$$

Atendendo à equação (3.14), a estimativa da incerteza na concentração da amostra apresenta três contribuições distintas provenientes das réplicas da amostra (σ_y^s, equação (3.16)), da imprecisão paramétrica referente à intercepção (b, equação (3.15)) e ao declive (m, equação (3.14)).

Através da expressão de propagação de erro indeterminado e independente e procedendo a algumas aproximações (por ex.: $\sigma_{\bar{y}_s} \approx \frac{\sigma_{fit}^2}{(m \times M)}$ chega-se à seguinte equação (3.16),

$$\sigma_{\bar{x}} = \frac{\sigma_{fit}}{m} \times \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{\sum_{i=1}^{n} (\bar{y}_s - \bar{y})^2}{m^2 \times \sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{\sigma_{fit}}{m} \times \sqrt{Q_1 + Q_2 + Q_3} \qquad (3.16)$$

onde o quociente Q1 está relacionado com a imprecisão das réplicas da amostra (M é o número de réplicas utilizadas), Q2 está relacionado com o erro na intercepção (N é o número de padrões da curva de calibração que foram utilizados no ajuste dos parâmetros da recta de calibração, devendo incluir o branco) e Q3 está relacionado com a imprecisão do declive.

b) Mínimos quadrados ponderados (WLS)

No caso de a variável dependente não ser homogénea (com variância estatisticamente distinta), é necessário proceder à pesagem estatística com a própria variância de cada ponto experimental. Visto que neste estimador, os elementos adquirem diferente importância na estimativa consoante o peso que lhe é atribuído.

Assim, a soma de quadrados dos resíduos do modelo é dada pela equação (3.17)

$$\chi^2 = \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \left(\frac{y_i - b_0 - m.x_i}{\sigma_i} \right)^2 = \sum_{i=1}^{n} w_i \times (y_i - b_0 - m.x_i)^2 \qquad (3.17)$$

Em que _i representa o desvio padrão da variável dependente. O desvio padrão do ajuste é dado pela equação (3.18)

$$\sigma_{w\,fit}^2 = \frac{\chi^2}{(n-p)} = \frac{\sum_{i=1}^{n} w_i \times (y_i - b_0 - m.x_i)^2}{(n-p)} \qquad (3.18)$$

e as respectivas equações normais

$$\begin{cases} \dfrac{\partial \chi^2}{\partial b_0} = 2 \times (\sum_{i=1}^{n} w_i \times (y_i - b_0 - m.x_i)) \times (-w_i) = 0 \\ \dfrac{\partial \chi^2}{\partial m} = 2 \times (\sum_{i=1}^{n} w_i \times (y_i - b_0 - m.x_i)) \times (-w_i \times x_i) = 0 \end{cases} \qquad (3.19)$$

que conduzem ao seguinte sistema de equações,

$$\left\{ A \mid ind_w \right\} = \left\{ \begin{array}{cc|c} W & S_{wx} & S_{wy} \\ S_{wx} & S_{wx^2} & S_{wxy} \end{array} \right\} \qquad (3.20)$$

onde

$W = \sum_{i=1}^{n} w_i$, $S_{wx} = \sum_{i=1}^{n} w_i \times x_i$, $S_{wy} = \sum_{i=1}^{n} wy_i$, $S_{wxy} = \sum_{i=1}^{n} w_i \times x_i \times y_i$, $S_{wx^2} = \sum_{i=1}^{n} w_i \times {x_i}^2$.

Este sistema de equações só possui solução se a matriz de projecto possuir determinante não nulo $\Delta = (N \times S_{x^2} - S_x \times S_x) = N \times \sum_{i=1}^{n} (x_i - \bar{x})^2$.

O declive (m) pode ser estimado através da regra de Cramer,

$$m = \frac{\Delta_{w1}}{\Delta_w} = \frac{N \times S_{wxy} - S_{wx} \times S_{wy}}{N \times S_{wx^2} - S_{wx} \times S_{wx}} = \frac{\sum_{i=1}^{n} w_i \times (x_i - \bar{x}) \times (y_i - \bar{y}).}{\sum_{i=1}^{n} w_i \times (x_i - \bar{x})^2.} \quad (3.21)$$

e a intercepção com o eixo das ordenadas através da equação do modelo que satisfaz o ponto central ($\bar{x}$; $\bar{y}$) da curva,

$$b = \bar{y} - m\bar{x} \quad (3.22)$$

Por sua vez, a imprecisão paramétrica é obtida através da matriz de covariância, em que $\sigma_b \, e \sigma_m$ podem ser obtidos a partir das equações (3.23) e (3.24)

$$\sigma_b = \frac{\sigma_{y_s}}{\sqrt{M}} = \sqrt{\frac{S_{wx^2}}{\Delta_w}} = \sqrt{\frac{\sum_{i=1}^{M} w_i \times {x_i}^2}{W \times \sum_{i=1}^{M} w_i \times (x_i - \bar{x}_w)^2}} \quad (3.23)$$

$$\sigma_m = \sqrt{\frac{W}{\Delta_w}} = \sqrt{\frac{1}{\sum_{i=1}^{M} w_i \times (x_i - \bar{x}_w)^2}} \quad (3.24)$$

em que Δ_w = W:Swx2 – Swx x Swx = W x $\sum_{i=1}^{M} w_i \times (x_i - \bar{x}_w)^2$ e $\bar{x}_w$ é a média ponderada dos valores da variável independente ($\bar{x}_w = \frac{\sum_{i=1}^{M} w_i \times x_i}{W}$).

Pode observar-se que a estimativa da imprecisão paramétrica no caso WLS, equações. (3.23) e (3.24), não recorre à variância do ajuste (estimativa da imprecisão residual) como no caso OLS, equações (3.13) e (3.14), respectivamente.

3.2.1.3 Diagnóstico da linearidade

É necessário recorrer a testes e critérios estatísticos que auxiliem a decidir qual o modelo mais adequado para proceder à calibração da resposta instrumental. Para tal pode-se recorrer ao teste de Mandel que se baseia em comparar o efeito da inclusão/exclusão de um parâmetro adicional na equação do modelo e verificar o seu efeito sobre a variância do ajuste.

Os modelos baseados em polinómios da variável independente com grau inteiro estão estatisticamente bem determinados quanto às estimativas paramétricas obtidas. Atendendo a este facto e ainda às características lineares da resposta instrumental, em geral, procura-se que a curva de calibração seja traduzida por um polinómio de baixo grau (de preferência 1 ou 2).

Regra geral, o teste de Mandel serve como teste de linearidade para verificar se a curva de calibração obtida pode ser representada por um polinómio de primeiro grau ou terá que se utilizar um polinómio de grau superior (em geral 2º grau).

3.2.1.4 Análise de outliers

Os valores experimentais que mais se afastam do modelo podem ser outliers. A regressão robusta pode ser útil para detectar eventuais outliers.

O teste de Mandel também pode ser utilizado para verificar se um determinado valor deve ou não ser excluído da curva de calibração por ser considerado um valor discrepante [outlier]. De igual modo compara-se o aumento na variância do ajuste ao introduzir o valor dúbio no conjunto dos pontos da curva de calibração com uma estimativa de erro puramente aleatório que resulta da variância do ajuste com (n - 1) valores experimentais

$$F = \frac{(SS_{(n-p)} - SS_{(n-p-1)}) \big/ (n-p)-(n-p-1)}{\sigma^2_{fit\,(n-p)-(n-p-1)}} \qquad (3.25)$$

Se este valor exceder o valor crítico $F^{u}_{0:01}$ (1;n-p-1), então o valor em causa afecta significativamente a variância do ajuste e este deve ser, por isso considerado um outlier.

3.2.2 Análise das amostras

3.2.2.1 Diagnóstico de valores discrepantes

Valores discrepantes são valores que não pertencem a uma determinada distribuição. Dado que as estimativas paramétricas são sensíveis a valores "contaminados", estes valores "outliers" produzem em geral erros de estimativa quer na posição (enviesamento) quer na dispersão (aumento da imprecisão). Como tal, para que este efeito de má estimativa seja evitado temos que testar inicialmente qualquer conjunto de dados antes de se proceder a qualquer estimativa.

Podem ser efectuados diversos testes estatísticos, sendo os mais significativos os que comparam a posição do valor duvidoso em relação à estimativa central. De entre os testes mais utilizados conta-se o teste de Grubbs (recomendado pela ISO).

A rejeição abusiva de valores experimentais de um conjunto de dados não constitui, em geral, um erro muito significativo no caso de o número de valores ser relativamente grande (m > 10). Como tal, o nível de significância utilizado em geral é de 5 % (α= 0:05).

Os valores discrepantes serão sempre extremos (inferior e/ou superior) que distam significativamente do corpo principal da distribuição. Como tal, o primeiro passo a dar é sempre o de ordenar os valores segundo a ordem crescente. Calcular as diferenças sucessivas entre os valores vizinhos auxilia a encontrar o valor dúbio - em geral é o extremo que se distanciar mais do valor mais próximo.

3.2.2.2 Estimativa da concentração

Assumindo que a curva de calibração é descrita através de um polinómio de primeiro grau (y =b + xi), as concentrações são estimadas através das equações (3.14) e (3.15), em que $\sigma_{\bar{x}}$ representa uma estimativa da incerteza na concentração da amostra estimada com base na curva de calibração. Com base nestes valores, o resultado a reportar é

$$\mu_A = x_A \pm t^{b}_{0.05(n-p)} \times \sigma_{x_A} \qquad (3.26)$$

O desempenho analítico do método é maximizado quando a imprecisão da estimativa da concentração da amostra é mínima ou seja, quando há:

- bom ajuste linear (σ_{fit} baixo);
- boa sensibilidade (declive da recta alto),

- grande número de padrões (n alto);
- um número razoável de réplicas da amostra ;
- leituras da amostra próximas do centróide ($\left(\overline{y_A} \approx \overline{y}\right)$;
- grande gama de concentrações de padrões $(\sum_{i=1}^{n}(x_i - \overline{x})^2 >> 0)$.

3.3 Análise multivariada dos dados

3.3.1 Pré-tratamento dos resultados

Para evitar trabalho desnecessário, a verificação da optimização deve estar constantemente presente durante este procedimento. Assim, deve-se verificar sistematicamente se os parâmetros estudados foram significativamente melhorados e se são adequados ao problema analítico inicial. Desta feita, é necessário efectuar um pré-tratamento dos resultados.

3.3.1.1 Valores omissos

Os valores omissos são um problema que muitas vezes dificulta a análise estatística dos resultados, pois eles condicionam as estimativas paramétricas e robustas, uma vez que os programas de tratamento estatístico como o MatLab, Octave dão problemas quando não são retirados estes valores. Assim sendo é necessário removê-los ou, em alguns casos, substituí-los pela média [80].

3.3.1.2 Estimativas paramétricas

A média aritmética, assim como o desvio-padrão que lhe está associado, são conceitos que geralmente oferecem poucas dúvidas, sendo calculados, apenas, em variáveis com a escala quantitativa.

Assim, média amostral ou simplesmente média, que se representa por $\overline{\mathbf{x}}$, é uma medida de localização do centro da amostra, e obtém-se a partir da equação (3.27),

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.27)$$

onde x_1, x_2, ..., x_n representam os elementos da amostra e **n** a sua dimensão.

O desvio padrão é uma medida de dispersão estatística, que mede a variabilidade dos valores à volta da média, pelo que, quando não existe variabilidade, tem 0 (o seu valor mínimo), indicando que todos os valores são iguais à média

A fórmula de cálculo do desvio padrão para os valores x_1, x_2, x_3,..., x_n de uma amostra é dada pela equação (3.28):

$$\bar{x} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_{n1} - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (3.28)$$

onde $\bar{x}$ é a média.

3.3.1.3 Estimativas robustas

As estimativas não paramétricas, também chamadas de Estimativas robustas, geralmente recorrem a operações simples tais como a ordenação não paramétrica dos valores e avaliação dos percentis. São exemplos destas estimativas a mediana e os intervalos de confiança com base em percentis.

Mediana é simplesmente o valor que se situa a meio da fila ordenada de valores, desde o mais baixo ao mais alto. Ou seja, ½ das amostras, da população ou distribuição de probabilidade terá valores inferiores ou iguais à mediana e ½ da população terá valores superiores ou iguais à mediana. Assim, tem que haver uma relação de ordem nos valores, pelo que a mediana pode ser calculada tanto para as variáveis ordinais como para as quantitativas puras.

A mediana pode ser calculada para um conjunto de observações ou para funções de distribuição de probabilidade. Se houver um número ímpar de valores ordenados é só verificar o valor que ocupa a posição central. Se houver um número par de valores ordenados soma-se os 2 valores que estão nas posições centrais e divide-se por 2.

Nalguns casos quer-se dividir a distribuição num número maior de partes. Quando se divide em 4 partes obtêm-se os quartis e a mediana tem o valor do segundo quartil:

0----------Q1----------Q2----------Q3----------Q4

Para indicar o grau de dispersão dos dados são usadas outro tipo de medidas de sumário, que se designam por medidas de dispersão, como o desvio padrão e o percentil, entre outras.

Um percentil é uma medida da posição relativa de uma unidade de observação em relação a todas as outras. O p-ésimo percentil tem no mínimo *p* % dos valores abaixo daquele ponto e no mínimo (100 - *p*) % dos valores acima.

Não se deve confundir percentis com percentagens**.** Um percentil é relacionado somente com a posição relativa de uma observação quando comparada com os outros valores.

3.3.2 Transformação de variáveis

As estimativas do valor central e da dispersão da distribuição permitem efectuar a transformação de variáveis. Sendo x uma variável aleatória com valor central μ_x e dispersão σ_x^2 ($x \sim D[\mu_x ; \sigma_x^2]$)

3.3.2.1 Centragem

A centragem **é** uma transformação simples que consiste em centrar cada variável em relação à sua média. A variável "e" corresponde à distribuição dos valores de x centrados em zero ($e \sim D[0; \sigma_x^2]$).

$$\mathrm{ei} = \mathrm{xi} - \mu_x \tag{3.29}$$

como consequência,

$$\mathrm{E[e]} = \mathrm{E[xi} - \mu_x \mathrm{]} = \; E[x] - E[\mu_x] = 0 \tag{3.30}$$

3.3.2.2 Escalamento

A variável r, que se obtém a partir da equação (3.31)

$$r_i = \frac{x_i}{\sigma_x} \tag{3.31}$$

diz-se escalada se a sua dispersão for unitária

$$r_i = \frac{x_i}{\sigma_x} \Rightarrow \mathrm{E[(r - E[r])^2]} = E\left[\left(\frac{\mathrm{x} - \mu_x}{\sigma_x^2}\right)^2\right] = \frac{\mathrm{E[(x} - \mu_x)^2]}{\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = 1 \tag{3.32}$$

Ao efectuar o escalamento ocorre uma alteração na posição, equação (3.33).

$$\mathrm{E}[\mathrm{r}]= E\left[\frac{x}{\sigma_x}\right]=\frac{1}{\sigma_x}\mathrm{E}[\mathrm{x}]=\frac{\mu_x}{\sigma_x}=1 \tag{3.33}$$

Razão pela qual este processo é geralmente acompanhado da centragem em zero (normalização).

3.3.2.3 Normalização

A normalização de distribuições é extremamente vantajosa dado que qualquer variável pode agora ser tratada de igual modo, independentemente das unidades em que deve ser considerada e da sua grandeza numérica. Esta operação permite ainda:

- a comparação directa de distribuições;
- converter qualquer distribuição numa distribuição estatística relevante (normalizada) e deste modo poder inferir conclusões estatísticas diversas;
- efectuar a simulação de distribuições baseadas em distribuições normalizadas, equação (3.34)

$$\mathrm{N}[\mu_x,\sigma^2]= \mu_x + \sigma^2 \times N[0.1]=1 \tag{3.34}$$

Normalização da variável z pode ser obtida pela equação (3.35)

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{3.35}$$

diz-se normalizada se se encontra escalada (dispersão unitária) e centrada em zero (z~D[0,1])

$$\begin{cases} E[z] = E\left[\dfrac{x_i - \mu_x}{\sigma_x}\right] = \dfrac{1}{\sigma_x}\mathrm{E}[x_i - \mu_x] = 0 \\ \mathrm{E}[(z_i - \mathrm{E}[z])^2] = E\left[\left(\dfrac{x - \mu_x}{\sigma_x^2}\right)^2\right] = \dfrac{1}{\sigma_x^2}\times\sigma_x^2 = 1 \end{cases} \tag{3.36}$$

A normalização de uma variável não altera a sua simetria nem a sua curtose, apenas permite centrar e normalizar a sua dispersão.

3.4 Análise de Componentes Principais

3.4.1 Cálculo do PCA usando método de covariância

O objectivo é transformar um conjunto de dados fornecido, X, de dimensão M, num conjunto de dados Y de dimensão L. Equivalentemente podemos encontrar a matriz Y, onde Y é a transformação de Karhunen-Loeve da matriz X [64], equação (3.37).

$$Y = KLT[X] \tag{3.37}$$

3.4.2 Organização de um conjunto de dados

Com o intuito de facilitar a compreensão, vamos supor uma situação prática, onde se tem um conjunto de dados, que contêm um agregado de observações de variáveis M. Neste caso pretende-se reduzir os dados de cada variável, de forma a que cada uma possa ser descrita por L, onde L<M. Acontece que os dados estão organizados como um conjunto N de vectores de dados, X_1, …. , X_N em que cada X_N representa um único grupo de observações de variáveis M. Para resolver este problema, é necessário começar por remover todas as variáveis nulas e informações irrelevantes para os cálculos, construindo, a partir daí, uma matriz X de dimensão MxN, com vectores coluna X_1, …., X_N, cada um deles com linhas M.

3.4.2.1 Cálculo da media empírica

Para calcular a média empírica é necessário encontrá-la ao longo de cada dimensão m = 1, …, M, ou seja, a dimensão correspondente à média de cada linha. Estes valores são colocados num vector empírico da média, “u”, de dimensão Mx1.

$$u = \frac{1}{N}\sum_{n=1}^{N} X[m,n] \tag{3.38}$$

3.4.2.2 Cálculo de desvios a partir da média

Neste caso, a subtracção é uma operação imprescindível porque permite a centragem dos dados, necessária para encontrar o componente que minimiza o erro da média da raiz quadrada dos dados aproximados [70].

Deste modo, começa-se por subtrair o vector da média empírica a cada coluna da matriz X, sendo os valores resultantes armazenados na matriz B (MxN).

$$B = X - uh \tag{3.39}$$

onde "h" é um vector 1xN linhas de todos os 1´s h[n]=1, para n=1, …, N.

De seguida calcula-se a matriz de covariância, que contém na sua diagonal principal as variâncias das colunas e nos restantes elementos as covariâncias.

3.4.2.3 Cálculo da matriz de covariância

A matriz empírica de covariância, C, pode ser determinada a partir do produto externo da matriz B com ela própria, equação (3.40).

$$C = [B \otimes B] = E\,[B.B^*] \tag{3.40}$$

Onde, E é o agente do valor esperado, $\otimes$ é o agente do produto externo e * é o agente de transposição conjugada. No entanto, se B for formado inteiramente por números reais, como é o caso de muitas aplicações, o "transporte conjugado" é o mesmo que o transporte regular.

A matriz de covariância contém os valores para a covariância entre todas as dimensões de um espaço:

$$C_{n\times n} = (C_{i,j}, C_{i,j} = cov\,(Dim_i, Dim_i)) \tag{3.41}$$

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix} \tag{3.42}$$

Os valores da diagonal superior descrevem a variância da respectiva dimensão. Uma vez que cov(x,y)=cov(y,x), a matriz é simétrica em relação à diagonal principal.

3.4.2.4 - Cálculo de vectores e valores próprios

Os vectores próprios de uma matriz correspondem às direcções que não são alteradas através da multiplicação (transformação) da matriz. Contudo nem todas as matrizes têm vectores próprios, sendo necessário que estas sejam quadradas (nxn).

Caso se escale um vector próprio antes da multiplicação obtém-se o mesmo múltiplo no resultado:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 24 \\ 16 \end{pmatrix} = 4 \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} \quad (3.43)$$

A noção de valor próprio está relacionada com a de vector próprio. Repare-se que na equação (3.43), independentemente do comprimento do vector próprio, o inteiro múltiplo do vector obtido após a sua multiplicação com a matriz mantinha-se. Nesse caso, esse valor (4) é o valor próprio correspondente ao vector, todavia, nem todas as matrizes possuem valores próprios [81]. Pode dizer-se que os valores e vectores próprios estão sempre associados.

Uma propriedade, "especialmente importante" para a utilização no PCA, está relacionada com o facto dos vectores próprios de uma matriz serem perpendiculares entre si, isto é, formam bases ortogonais.

Para efeitos matemáticos é usual determinar os vectores - próprios na sua versão normalizada, isto é, com comprimento 1.

Para se encontrar os vectores e valores próprios, tem de calcular a matriz V, que diagonaliza a matriz de covariância C. Enquanto, a matriz D é a matriz diagonal dos valores próprios da matriz de covariância C, equação (3.44).

$$V^{-1} \times C \times V = D \quad (3.44)$$

Este processo envolve o uso de algoritmos que se baseiam em operações que calculam os vectores e valores pretendidos.

Os algoritmos podem encontram-se disponíveis como sub componentes da maioria dos sistemas de calculo de matrizes algébricas como é o caso do MatLab [82], Mathematica [83], GNU Octave [84] e SciPy [85].

3.5 Modelação multivariada

3.5.1 Regressão de Componentes Principais: PCR

Os valores X-scores são escolhidos para explicar, tanto quanto possível, do factor de variação. Esta aproximação concede direcções informativas quanto ao factor do espaço, mas não deve ser associada com a forma da superfície prevista [72].

3.5.2 Análise da Redundância Máxima

Os valores Y-scores são escolhidos para explicar, tanto quanto possível, a variação prevista de Y. Esta aproximação procura direcções no factor espaço, associadas com a maior variação nas respostas, não se apresentando as previsões muito precisas [72].

A regressão PLS decompõe tanto X como Y em produtos de um conjunto de factores ortogonais, como um conjunto específico de carregamentos.

$X=TP'$

com $T' T=I$

Y estima-se como sendo

$Y=TBC'$

Onde B é a matriz diagonal com "repressão de pesos" como elementos diagonais e C é a"matriz de peso" de variáveis dependentes. As colunas de T são vectores latentes, que podem ser escolhidos de diferentes formas: qualquer conjunto de vectores ortogonais spanning (de mediação) a coluna do espaço X pode ser usado para desempenhar um papel de T. Contudo, para especificar T são necessárias condições adicionais. É preciso encontrar dois conjuntos de pesos w e de forma a criar colunas x e y para que a sua covariância seja máxima.

O objectivo é obter a primeira parte de vectores $t=Xw$ e $u=Y$ e com a repressão que $w'w=1$, $t't=1$ e $t'u$ seja máxima. Quando o primeiro vector latente é encontrado, é subtraído de X e Y, repetindo-se o processo até que X se torne numa matriz nula.

O primeiro passo é criar duas matrizes: $E=X$ e $F=Y$, que serão centradas nas colunas e normalizadas, isto é, transformadas em resultados Z. A soma dos quadrados das duas matrizes é conotada de SSx e SS_Y. Antes de iniciar com a repetição, o vector U começa com valores aleatórios.

Capítulo 4

4.1 Pré-tratamento dos dados

Antes de iniciar o tratamento dos resultados convém garantir que estes mesmos não serão responsáveis por inviabilizar o cálculo ou o algoritmo através da introdução de algum tipo de singularidade numérica. Para tal, é necessário estabelecer antecipadamente objectivos e estratégias para minimizar insucessos de cálculo e maximizar o poder discriminante dos dados.

A matriz original de dados refere-se a teores analíticos de diferentes analitos (variáveis) em amostras de sedimentos de corrente (objectos).

A matriz inicial dos dados é composta por 26 variáveis (Fe, Ba, P, Cu, Cr, Ag, B, Zn, Sb, Sb-Colorimetria, Pb, Sn, Sn- Colorimetria, Ni, V, Mn, Be, Mo, As, W, W- Colorimetria, Co, Y, Cd, Nb, U), organizadas sob a forma de colunas, e com 684 de amostras (objectos), organizados por linhas. Verificou-se que alguns destes objectos apresentavam omissões de valores.

4.1.1 - Remoção de valores omissos

No sentido de evitar erros na leitura de dados pelo Octave[5], foi necessário remover os objectos (linhas) que continham valores omissos. Após esta triagem, ficou-se com 516 objectos, com os quais se trabalhou nas etapas seguintes.

4.1.2 - Diagnóstico das variáveis

Em primeiro lugar, antes de se realizar qualquer tipo de transformação das variáveis é conveniente conhecer, em cada caso **a)** o perfil da distribuição de cada variável e **b)** verificar se estão presentes outliers, para que possam ser correctamente inferidas as respectivas estimativas central (Xm) e de dispersão (sx). Convém, ainda, verificar se as escalas utilizadas são ou não da mesma ordem de grandeza para que possam ser utilizadas dimensões aceitáveis, para que todas as variáveis podem ser contabilizadas da mesma forma.

Na tabela 4.1 estão presentes as seguintes estimativas: estimativa paramétrica da média (Xm), desvio padrão (sx), variância (Var), estimativa de simetria (Skew), curtose (Kurt), teste combinado de simetria e curtose (T_{SK}) e valor da probabilidade correspondente para que a distribuição possa ser considerada normal (p[Norm]).

[5]O Octave é um programa vocacionado para cálculo matricial, compatível com o MatLab, mas com distribuição livre e acessível a todos.

Tabela 4.1 – Resumo das características das variáveis e sua eventual caracterização como distribuição normal (n = 516).

Variável	**Fe**	**Ba**	**P**	**Cu**	**Cr**	**Be**	**U**	**U**	**Cd**
Xm	3.63	385.48	543.65	30.33	240.12	1.44	1.06	1.06	0.53
Sx	0.82	87.06	166.67	11.06	77.28	0.540	0.504	0.504	0.226
Var	6.70E-01	7.58E+03	2.78E+04	1.22E+02	5.97E+03	2.90E-01	2.50E-01	2.50E-01	5.00E-02
Skew	-0.46	0.24	0.99	1.15	0.99	0.52	1.97	1.97	15.43
Kurt	3.35	3.28	4.52	9.81	4.00	2.08	8.95	8.95	298.72
T_{SK}	2.13E+01	6.97E+00	1.37E+02	1.14E+03	1.08E+02	4.24E+01	1.12E+03	1.12E+03	1.96E+06
p[Norm]	0.000	0.031	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Variável	**Nb**	**Ag**	**B**	**Zn**	**Sb**	**Sb-Color.**	**Pb**	**Sn**	**Sn- Color.**
Xm	5.01	0.12	27.64	71.12	11.10	5.40	21.46	5.82	0.56
Sx	0.220	0.06	10.76	33.04	16.937	37.552	12.830	2.100	0.656
Var	0.05	0.00	115.70	1091.88	286.86	1410.12	164.60	4.41	0.43
Skew	22.72	3.81	0.63	4.07	21.12	21.49	5.45	2.87	11.14
Kurt	519.00	19.11	4.48	38.44	467.04	479.59	58.87	13.42	128.05
T_{SK}	5.94E+06	7.00E+03	8.28E+01	2.92E+04	4.80E+06	5.07E+06	7.17E+04	3.12E+03	3.57E+05
p[Norm]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Variável	**Ni**	**V**	**Mn**	**Mo**	**As**	**W**	**W- Color.**	**Co**	**Y**
Xm	33.65	106.05	283.91	1.40	15.85	5.94	1.82	7.20	8.38
Sx	9.799	31.274	163.194	0.949	11.025	6.102	4.229	4.355	3.005
Var	9.60E+01	9.78E+02	2.66E+04	9.00E-01	1.22E+02	3.72E+01	1.79E+01	1.90E+01	9.03E+00
Skew	0.61	-0.64	3.03	2.69	3.82	18.21	17.80	2.84	0.90
Kurt	7.25	2.65	20.75	10.01	31.53	381.46	371.45	16.82	3.79
T_{SK}	4.32E+02	3.80E+01	7.77E+03	1.71E+03	1.93E+04	3.20E+06	3.03E+06	4.93E+03	8.44E+01
p[Norm]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Na análise realizada à tabela 4.1 verificou-se que nenhuma das variáveis apresenta um comportamento normal, sendo que apenas a segunda variável (Ba) tem 3.1 % de hipóteses de poder ser considerada normal.

Perante esta situação, é necessário estar consciente de que estas anomalias de não-normalidade irão condicionar as estimativas paramétricas central (média) e de dispersão (desvio padrão), especialmente quando se pretender centrar e/ou normalizar as variáveis.

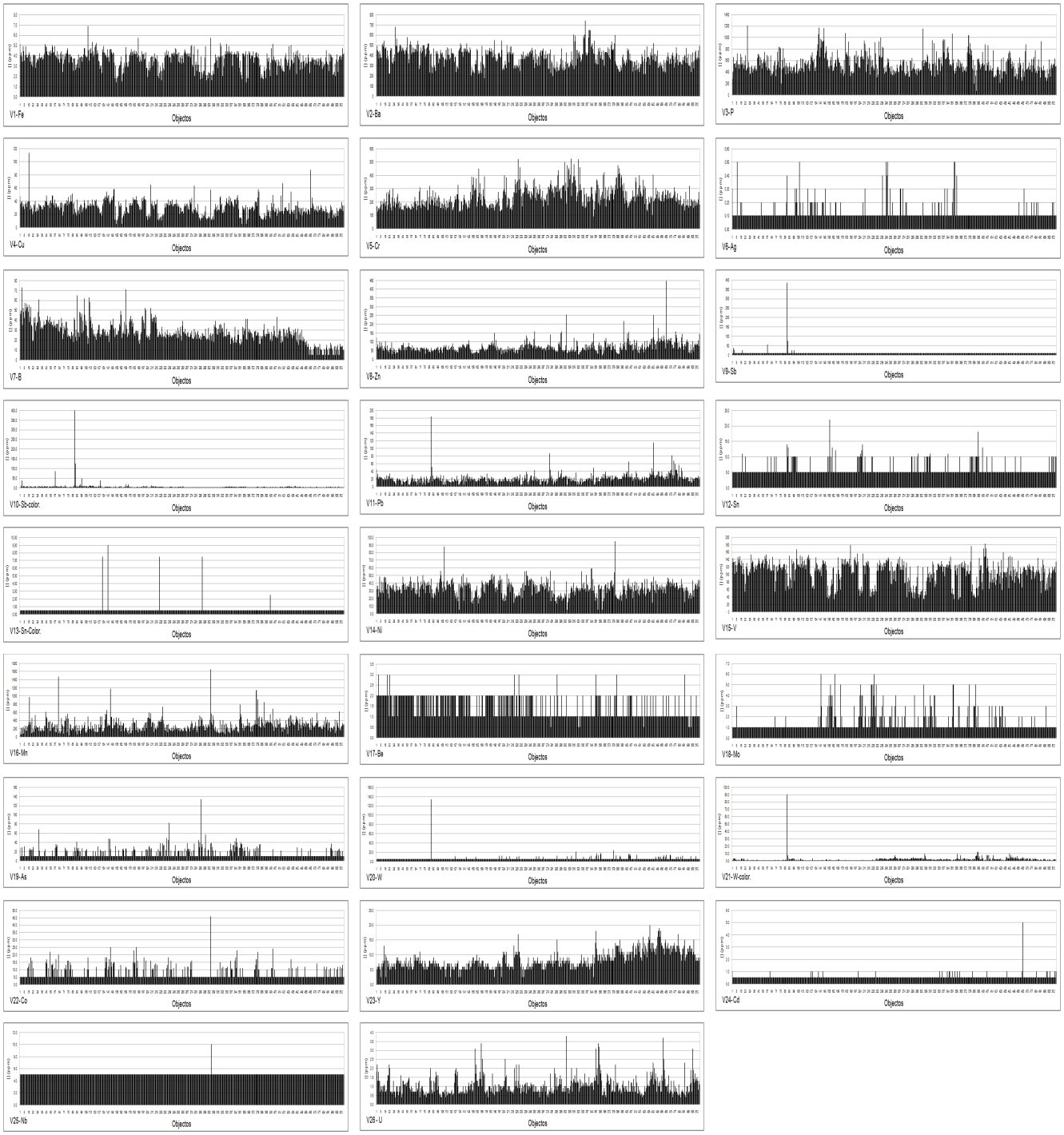


Figura 4.1 - Representação gráfica das 26 variáveis em estudo. NOTA: estas figuras constam do anexo A2, como a figura A2.1, caso deseje ver mais em detalhe.

Da figura 4.1 verifica-se que 12 destas variáveis (V06 (Ag), V09 (Sb), V10 (Sb-Color.), V12 (Sn), V13 (Sn-Color.), V17 (Be), V18 (Mo), V19 (As), V20 (W), V22 (Co), V24 (Cd) e V25 (Nb)) apresentam baixa capacidade discriminante associada a severos casos de outliers.

Estas anomalias podem ter forte impacto sobre as estimativas paramétricas no pré-processamento dos dados. No sentido de minimizar os efeitos de falsa estimativa devida aos valores discrepantes de cada distribuição realizou-se um estudo de estimativa da dispersão robusta com base nos pressentis da distribuição. Estes resultados estão sistematizados na tabela 4.2.

Tabela 4.2 – Estimativas robustas de dispersão (baseadas em percentis, $S_{(1.0)}$ a $S_{(3.0)}$) e central (mediana, $X_{0.50}$).

Analito	**Fe**	**Ba**	**P**	**Cu**	**Cr**	**Ag**	**B**	**Zn**	**Sb**
Variável	**V01**	**V02**	**V03**	**V04**	**V05**	**V06**	**V07**	**V08**	**V09**
$S_{(1.0)}$	0.950	99.185	178.685	11.843	83.500	0.050	11.843	28.500	2.500
$S_{(1.5)}$	0.867	83.683	164.175	10.333	77.175	0.033	10.667	26.017	1.842
$S_{(2.0)}$	0.777	79.356	159.143	9.250	71.606	0.050	11.750	27.750	2.250
$S_{(2.5)}$	0.720	80.266	164.311	10.092	70.057	0.080	11.273	27.387	7.100
$S_{(3.0)}$	0.716	80.163	163.592	13.567	70.188	0.067	10.970	38.147	20.707
Xm	3.628	385.481	543.649	30.333	240.124	0.121	27.640	71.122	5.397
$X_{0.50}$	3.800	389.000	511.500	31.000	225.000	0.100	26.000	65.000	2.500
Sx	0.811	84.861	166.113	11.122	74.686	0.058	11.309	29.881	9.939

Analito	**Sb-Color.**	**Pb**	**Sn**	**Sn-Color.**	**Ni**	**V**	**Mn**	**Ni**	**V**
Variável	**V10**	**V11**	**V12**	**V13**	**V14**	**V15**	**V16**	**V14**	**V15**
$S_{(1.0)}$	0.000	9.000	2.500	0.000	10.843	39.843	150.870	10.843	39.843
$S_{(1.5)}$	0.000	8.667	1.667	0.000	9.333	31.683	125.033	9.333	31.683
$S_{(2.0)}$	0.000	9.524	1.274	0.000	8.750	27.750	123.356	8.750	27.750
$S_{(2.5)}$	3.092	12.146	1.600	0.146	8.600	26.546	172.634	8.600	26.546
$S_{(3.0)}$	10.572	18.350	2.147	1.167	13.688	24.157	234.515	13.688	24.157
Xm	11.105	21.455	5.816	0.561	33.647	106.047	283.909	33.647	106.047
$X_{0.50}$	10.000	20.000	5.000	0.500	34.000	113.000	260.500	34.000	113.000
sx	4.926	12.092	1.888	0.526	10.417	30.494	166.380	10.417	30.494

Analito	**Mn**	**Be**	**Mo**	**As**	**W**	**W-Color.**	**Co**	**Y**	**Cd**	**Nb**	**U**
Variável	**V16**	**V17**	**V18**	**V19**	**V20**	**V21**	**V22**	**V23**	**V24**	**V25**	**V26**
$S_{(1.0)}$	150.870	0.500	0.500	8.500	0.000	1.350	3.500	3.500	0.000	0.000	0.500
$S_{(1.5)}$	125.031	0.333	0.667	7.842	1.667	1.100	3.333	2.667	0.000	0.000	0.418
$S_{(2.0)}$	123.350	0.250	1.000	7.500	2.000	1.800	3.250	2.500	0.125	0.000	0.450
$S_{(2.5)}$	172.622	0.500	0.800	7.800	2.273	1.740	3.600	3.100	0.100	0.000	0.547
$S_{(3.0)}$	234.511	0.417	0.833	11.930	3.152	1.950	3.333	2.745	0.083	0.000	0.549
Xm	283.910	1.437	1.403	15.849	5.938	1.823	7.203	8.385	0.535	5.010	1.055
$X_{0.50}$	260.500	1.000	1.000	10.000	5.000	1.500	5.000	8.000	0.500	5.000	1.000
sx	166.310	0.412	0.778	8.867	2.092	1.619	3.406	2.924	0.081	0.000	0.495

Da tabela 4.2 confirma-se que mais de 99.6% dos valores contidos em V25 (Nb) são constantes; cerca de 95% da informação contida nas variáveis V10 (Sb-Color.) e V13 (Sn-Color.) é constante e cerca de 87% da informação contida em V24 (Cd) é, também, constante. Já a variável V20 (W) contém cerca de 75.8% de informação constante.

Daqui se conclui que as variáveis menos relevantes para este estudo discriminante são, por ordem crescente de relevância, V25 (Nb), V10 (Sb-Color.), V13 (Sn-Color.), V24 (Cd) e V20 (W). Verifica-se ainda que as variáveis V06 (Ag), V10 (Sb-Color.), V12 (Sn), V17 (Be), V18 (Mo), V19 (As), V22 (Co) e V23 (Y) apresentam severos casos de outliers.

Para averiguar se estes valores são ou não outliers, pode-se fazer uma comparação entre o valor considerado outlier e a sua média. Para tal operação utiliza-se o teste t-student, em que as hipóteses a considerar vão no sentido de o valor em causa ($x_?$) pertencer ou não à distribuição. Deste modo, a hipótese

nula (H0) considera que o valor pertence à distribuição e a hipótese alternativa (H1) que o valor não pertence à distribuição.

A função discriminante a estimar é dada pela equação 4.1.

$$\mathrm{TV} = \frac{(x_? - \bar{X})}{s_{\bar{X}}} = \frac{(x_? - \bar{X})}{s_x/\sqrt{n}} \tag{4.1}$$

em que $\bar{X}$ e $S_{\bar{x}}$ são estimativas populacionais, obtidas com a exclusão do ponto em questão, porque a média ($\bar{X}$) e o desvio padrão ($S_{\bar{x}}$) são estimativas paramétricas sensíveis a outliers. Como tal, têm de ser obtidas por remoção do valor duvidoso da distribuição em causa, sob pena de poderem estar enviesadas por acção do valor deste.

O valor calculado para a função discriminante (TV) deve ser confrontado com o respectivo valor da distribuição t-student referente a ν = (n - 2) graus de liberdade e para o nível de confiança de 99% (teste bilateral), tabela A1.2 do anexo A1. Como simplificação, pode-se assumir que, quando uma variável possui mais que um valor extremo com características de outlier, testa-se primeiro o valor do "eventual outlier" mais próximo da distribuição. Caso este seja outlier, os restantes, mais afastados do centro, também o serão.

Utilizando este método de filtragem de valores discrepantes, foram detectados e testados 22 eventuais outliers, os quais resultaram na remoção de 15 valores extremos. Deste modo, os objectos passaram a ser 501.

Quanto às variáveis, duas delas deixaram de ser relevantes, já que agora adquirem sempre valores constantes (variáveis V13 (Sn-Color.) e V25 (Nb)). As restantes variáveis passaram a assumir distribuições mais uniformes em termos de dispersão de valores, ver figura 4.2.

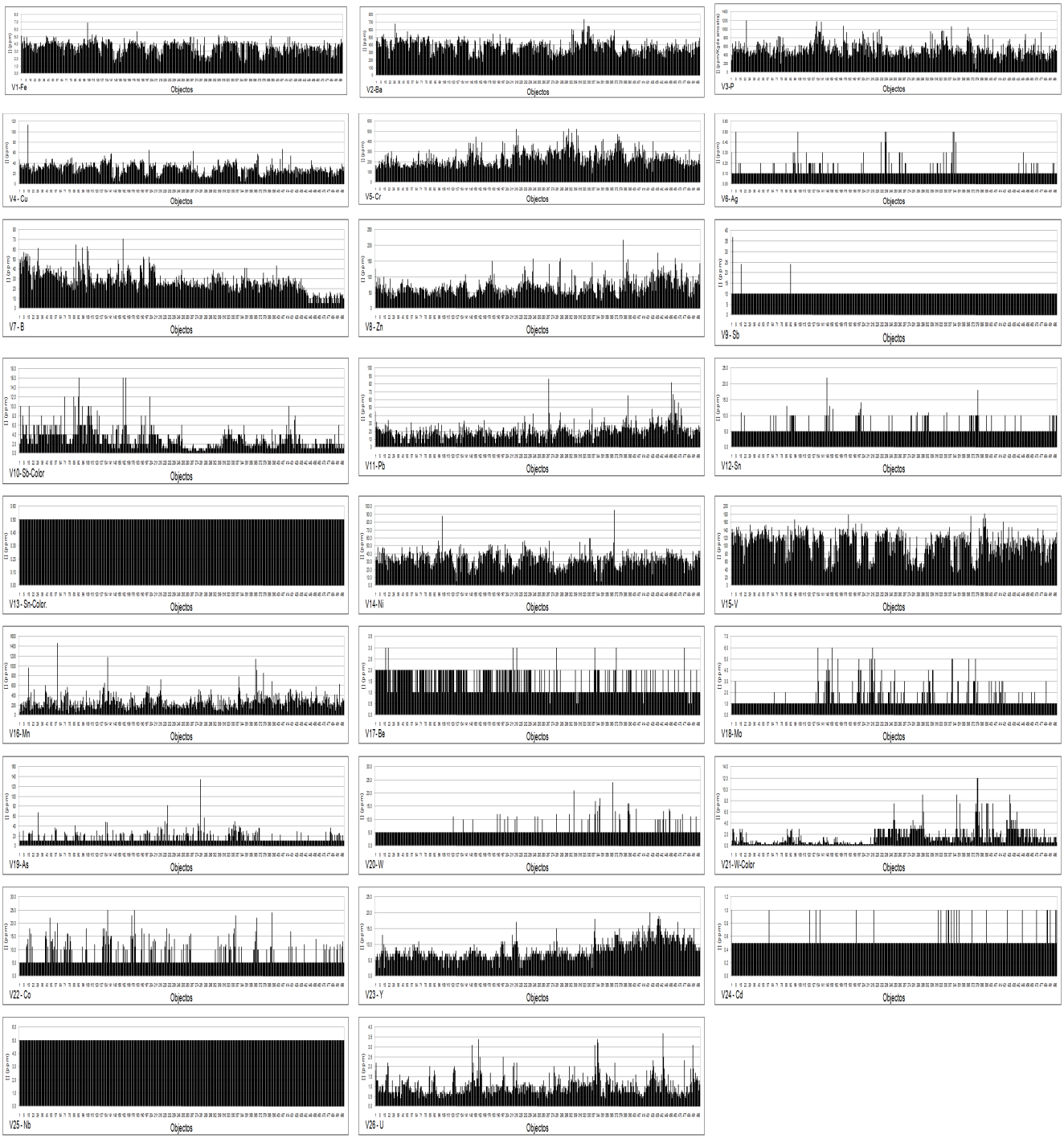


Figura 4.2 - Representação das variáveis em estudo após terem sido removidos os 15 outliers (n = 501). No anexo A2, como Figura A2.2, encontram-se as mesmas figuras com maior detalhe.

Pela figura 4.2 pode verificar-se que existem variáveis (Be, Sb, Sn-color, W, Cd e Nb) que não possuem grande informação para a análise do sistema Geo-Químico que se pretende efectuar, pelo que podem ser removidas deste conjunto.

Na tabela 4.3 encontram-se sistematizadas as estimativas robustas e paramétricas de cada uma das variáveis após a remoção dos valores anómalos.

Tabela 4.3- Estimativas robustas de dispersão (baseadas em percentis, $S_{(1.0)}$ a $S_{(3.0)}$) e central (mediana, $X_{0.50}$).

Analito	**Fe**	**Ba**	**P**	**Cu**	**Cr**	**Ag**	**B**	**Zn**	**Sb**
Variável	**V01**	**V02**	**V03**	**V04**	**V05**	**V06**	**V07**	**V08**	**V09**
$S_{(1.0)}$	0.950	99.185	178.685	11.843	83.500	0.050	11.843	28.500	2.500
$S_{(1.5)}$	0.867	83.683	164.175	10.333	77.175	0.033	10.667	26.017	1.842
$S_{(2.0)}$	0.777	79.356	159.143	9.250	71.606	0.050	11.750	27.750	2.250
$S_{(2.5)}$	0.720	80.266	164.311	10.092	70.057	0.080	11.273	27.387	7.100
$S_{(3.0)}$	0.718	80.165	163.480	13.120	70.102	0.071	10.981	38.102	20.707
Xm	3.634	385.481	543.649	30.333	240.124	0.121	27.640	71.122	5.397
$X_{0.50}$	3.800	389.000	511.500	31.000	225.000	0.100	26.000	65.000	2.500
Sx	0.809	84.861	166.113	11.122	74.686	0.058	11.309	29.881	9.939

Analito	**Sb-Color.**	**Pb**	**Sn**	**Sn-Color.**	**Ni**	**V**	**Mn**	**Ni**	**V**
Variável	**V10**	**V11**	**V12**	**V13**	**V14**	**V15**	**V16**	**V14**	**V15**
$S_{(1.0)}$	0.000	9.000	2.500	0.000	10.843	39.843	150.870	10.843	39.843
$S_{(1.5)}$	0.000	8.667	1.667	0.000	9.333	31.683	125.033	9.333	31.683
$S_{(2.0)}$	0.000	9.524	1.274	0.000	8.750	27.750	123.356	8.750	27.750
$S_{(2.5)}$	3.092	12.146	1.600	0.000	8.600	26.546	172.634	8.600	26.546
$S_{(3.0)}$	10.551	18.350	2.147	0.000	13.688	24.157	234.515	13.688	24.157
Xm	11.105	21.455	5.816	0.500	33.647	106.047	283.909	33.647	106.047
$X_{0.50}$	10.000	20.000	5.000	0.500	34.000	113.000	260.500	34.000	113.000
Sx	4.926	12.092	1.888	0.000	10.417	30.494	166.380	10.417	30.494

Analito	**Mn**	**Be**	**Mo**	**As**	**W**	**W-Color.**	**Co**	**Y**	**Cd**	**Nb**	**U**
Variável	**V16**	**V17**	**V18**	**V19**	**V20**	**V21**	**V22**	**V23**	**V24**	**V25**	**V26**
$S_{(1.0)}$	150.870	0.500	0.500	8.500	0.000	1.350	3.500	3.500	0.000	0.000	0.500
$S_{(1.5)}$	125.033	0.333	0.667	7.842	1.667	1.100	3.333	2.667	0.000	0.000	0.418
$S_{(2.0)}$	123.356	0.250	1.000	7.500	2.000	1.800	3.250	2.500	0.125	0.000	0.450
$S_{(2.5)}$	172.634	0.500	0.800	7.800	2.273	1.740	3.600	3.100	0.100	0.000	0.547
$S_{(3.0)}$	234.502	0.422	0.812	11.902	3.134	1.944	3.300	2.728	0.091	0.000	0.541
Xm	283.911	1.437	1.403	15.849	5.938	1.823	7.203	8.385	0.535	5.000	1.055
$X_{0.50}$	260.501	1.000	1.000	10.000	5.000	1.500	5.000	8.000	0.500	5.000	1.000
Sx	166.375	0.412	0.778	8.867	2.092	1.619	3.406	2.924	0.081	0.000	0.495

Comparando as tabelas 4.2 e 4.3 verifica-se que as estimativas robustas são praticamente inalteradas após a remoção de 15 valores em 516 valores iniciais (correspondendo à remoção de cerca de 2.9 % dos valores). Contudo, como seria de esperar, devido ao modo de cálculo, já as estimativas paramétricas das variáveis sofrem grandes variações com a presença desses valores discrepantes.

A presença deste conjunto de outliers implica, quanto à estimativa central, desvios sistemáticos significativos apenas em cinco variáveis: a variável V26 (U) é a que apresenta maior diferença (desvio de -79.0 %), seguida de V10 (Sb-Color., desvio de -23.0 %), V21 (W-Color., desvio de -8.6 %), V09 (Sb, desvio de -6.1 %) e V20 (W, desvio de -4.0 %). As restantes variáveis apenas apresentam desvios inferiores a 1.0 %.

Já quanto à estimativa da dispersão, há 12 variáveis com desvios superiores a 5.0 % e mais 3 com desvios superiores a 1.0 %: V10 (Sb-Color., desvio de 2414.7 %), V09 (Sb, desvio de 1034.2%), V11 (Pb, desvio de 453.4 %), V26 (U, 353.9 %), V22 (Co, 164.0 %), V24 (Cd, 92.5), V21 (W-Color., 84.0), V12 (Sn, 78.5

%), V20 (W, 44.4 %), V23 (Y, 25.7 %), V08 (Zn, 24.9 %), V16 (Mn, 8.0 %), V04 (Cu, 3.2 %), V18 (Mo, 2.5 %) e V06 (Ag, 1.5 %). Desta análise verifica-se que as estimativas paramétricas de dispersão são muito mais susceptíveis que as de posição, em relação à sua robustez com valores anormais.

Como resultado deste pré-processamento, tem-se agora dois tipos de matrizes com dados – o conjunto original (X(516×26), n = 516) e o conjunto filtrado (Xf(501×24), n = 501). Ambos os conjuntos serão estudados e os respectivos resultados discutidos de seguida.

4.2 Análise de componentes principais

Dado que a variância é um estimador paramétrico da dispersão dos dados de uma distribuição, e que esta depende da posição da respectiva distribuição, para se proceder à análise da variabilidade dos dados, através das suas componentes principais, é necessário que os valores sejam no mínimo centrados[6].

No caso de as variáveis possuírem dimensões muito distintas é ainda conveniente escalá-las de forma a trabalhar com todas as variáveis em igualdade de circunstâncias, sem serem afectadas por efeitos de escala. Neste caso, após a centragem, o escalamento faz com que estas passem a estar normalizadas[7].

A normalização de variáveis tem ainda o benefício de converter estas em grandezas dimensionais, com média nula e dispersão unitária. Deste modo, podem ser facilmente comparáveis com algumas das distribuições estatísticas mais convenientes.

Já que qualquer uma destas transformações depende de estimativas paramétricas (de posição e de dispersão) é, por isso, necessário avaliar a normalidade das distribuições para evitar anomalias de falta de normalidade.

Para se perceber o efeito destas transformações na análise PCA foram considerados diferentes processos de tratamentos:

P0 - sem qualquer transformação

P1 - centragem na média global

P2 - centragem na média de cada variável

P3 - normalização (Xm, sx)

P4 - normalização robusta (mediana e estimativa robusta da dispersão baseada em percentis)

P5 - normalização (Xm, sx) após remoção de 15 outliers (conjunto Xf(501×24), n = 501)

De seguida apresentam-se e discutem-se os resultados obtidos na análise PCA.

[6] A centragem corresponde a subtrair a cada valor da variável em questão o seu valor central, regra geral a média.

[7] Uma variável diz-se normalizada quando as estimativas de posição e dispersão são 0 e 1, respectivamente.

4.2.1 Número de componentes principais

O número de componentes principais de um sistema não é sempre uma grandeza bem definida, servindo esta para simplificar a abordagem ao sistema multivariado.

Existem três critérios que permitem definir o número de componentes principais (p) do sistema que são conhecidos por critérios de Pearson, de Kaiser e "Scree plot".

O critério de Pearson, também conhecido como "regra da recuperação de 80 %", define o número de componentes principais que caracterizam o sistema como sendo o número mínimo de componentes mais relevantes[8], permitindo reproduzir ou descrever cerca de 4/5 da variabilidade total.

O critério de Kaiser limita-se a definir as componentes principais com base no respectivo valor relativo – o número de componentes necessárias à descrição do sistema em causa coincide com o número de valores próprios que são superiores ao respectivo valor médio.

O scree plot – gráfico representativo dos valores próprios, ordenados por ordem decrescente de relevância em função do índice da componente – permite estabelecer a dimensão p através do número de componentes que apresentam contribuição acima da tendência basal.

Na tabela 4.3 encontram-se resumidos os resultados obtidos com os diferentes processos de acondicionamento dos dados originais (P0 a P5).

[8] As componentes são ordenadas por relevância com base no valor próprio associado.

Tabela 4.4 – Resultados da análise do número de componentes principais para os 6 processamentos propostos em 4.2.

#	P0	VarX%	CumVarX%
1	631538.21	93.39	93.39
2	21282.97	3.15	96.54
3	10846.60	1.60	98.14
4	8869.54	1.31	99.45
5	1788.42	0.26	99.72
6	1201.69	0.18	99.89
7	329.89	0.05	99.94
8	114.31	0.02	99.96
9	97.24	0.01	99.97
10	77.04	0.01	99.99

#	P1	VarX%	CumVarX%
1	500638.75	91.23	91.23
2	22695.55	4.14	95.36
3	11680.57	2.13	97.49
4	8896.69	1.62	99.11
5	2024.51	0.37	99.48
6	1156.36	0.21	99.69
7	1028.32	0.19	99.88
8	300.47	0.05	99.93
9	106.02	0.02	99.95
10	97.17	0.02	99.97

#	P2	VarX%	CumVarX%
1	39286.82	54.23	54.23
2	17398.14	24.02	78.25
3	8906.31	12.29	90.55
4	3263.44	4.51	95.05
5	1787.69	2.47	97.52
6	1113.32	1.54	99.06
7	304.14	0.42	99.48
8	113.79	0.16	99.63
9	97.24	0.13	99.77
10	73.34	0.10	99.87

#	P3	VarX%	CumVarX%
1	5.53	21.26	21.26
2	4.29	16.51	37.78
3	2.25	8.66	46.44
4	1.92	7.37	53.81
5	1.26	4.84	58.65
6	1.16	4.45	63.10
7	1.08	4.15	67.25
8	0.99	3.83	71.07
9	0.89	3.44	74.51
10	0.86	3.29	77.81

#	P4	VarX%	CumVarX%
1	237.86	95.79	95.79
2	2.28	0.92	96.71
3	1.70	0.68	97.39
4	1.60	0.64	98.03
5	1.09	0.44	98.47
6	0.87	0.35	98.82
7	0.45	0.18	99.00
8	0.42	0.17	99.17
9	0.33	0.13	99.31
10	0.29	0.12	99.43

#	P5	VarX%	CumVarX%
1	5.71	23,78	23.78
2	2.48	10,34	34.12
3	2.07	8,62	42.74
4	1.47	6,11	48.85
5	1.21	5,06	53.91
6	1.12	4,67	58.58
7	1.05	4,37	62.95
8	1.00	4,17	67.12
9	0.89	3,71	70.83
10	0.83	3,45	74,28

- indica da componente estimada; VarX% - variabilidade descrita pela componente principal em questão; CumVarX% - variabilidade cumulativa justificada.

Olhando para os resultados da tabela 4.4, de acordo com o critério de Pearson (recuperação de 80% variabilidade), o número de componentes principais seria p = 1 (P0, P1, P4), p = 3 (P2, centragem por variável) e p = 11 (P3, normalização). Já no caso P5 (conjunto filtrado) p assume um valor superior a 10.

Como seria esperado, a centragem e escalamento das variáveis faz com que se consiga penetrar mais em todos os domínios da informação contida nos dados levando a que seja necessário aumentar o número de componentes principais para conseguir descrever a mesma quantidade de informação.

Atendendo agora ao critério de Kaiser (valores próprios mais significativos), o número de componentes principais seria p = 1 (P0 - sem tratamento, P4 - com normalização robusta), p = 2 (P1 - centragem global), p = 4 (P2 - centragem por variável) e p = 7 (P3 - normalização e P5 - normalização com dados isentos de outliers).

Esta aproximação revela, uma vez mais, que a normalização paramétrica permite extrair mais informação dos dados já que é necessário um maior número de componentes para descrever o sistema. Destas duas abordagens verifica-se, ainda, que a normalização robusta não funciona bem – provavelmente porque as estimativas robustas são pouco sensíveis a outliers e, deste modo, mantém a informação "descompactada"[9] e não normalizada.

Na figura 4.3 encontram-se resumidos os diferentes "scree plots".

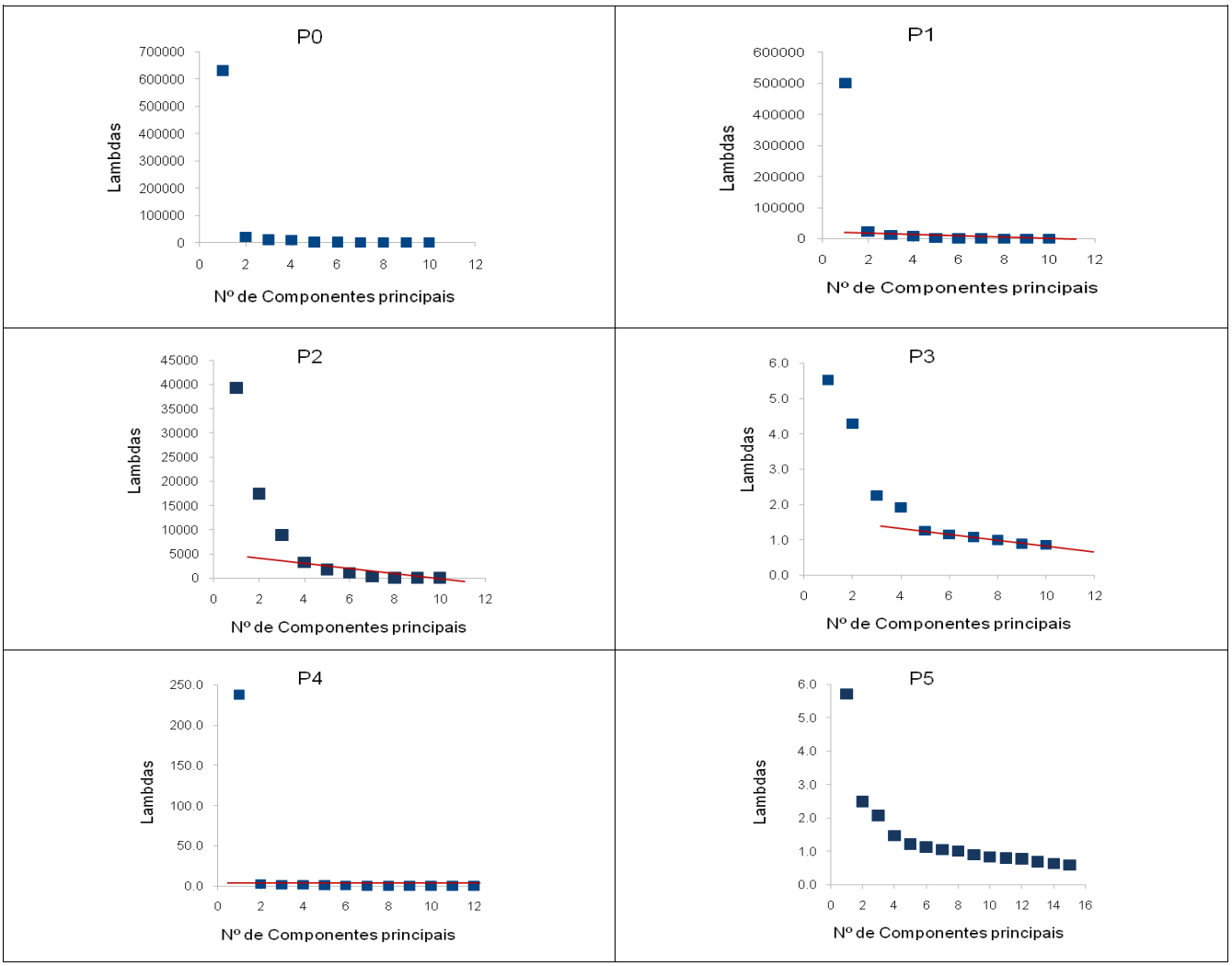


Figura 4.3 – Representação dos valores próprios em função do número de factores considerados (scree plot).

[9] Os estimadores paramétricos para a dispersão sofrem uma inflação no seu valor aquando da presença de valores discrepantes na amostra e, deste modo, permitem a "compressão" da informação contida na variável tendendo para convertê-la numa variável normal. Já o respectivo estimador robusto não se revela muito influenciado e, portanto, não comprime tanto esta informação.

De acordo com o critério "Scree plot" (valores acima da contribuição basal), o número de componentes principais seria p = 1 (P0, P1, P4), p = 3 (P2 - centragem por variável) e p = 4 (P3 - normalização e P5 - normalização e ausência de outliers).

Assim, resumindo a análise em termos das componentes principais que descrevem o sistema, teremos:

a) p = 1 no caso do tratamento P0 (sem qualquer transformação) e P4 (normalização robusta) – estes pré-tratamentos não permitem evidenciar a informação residual contida nas variáveis sendo a variabilidade dada essencialmente pela posição (no primeiro caso) e pela dispersão (no segundo caso);

b) p = 1 ou 2 no caso P1 (centragem global) – a centragem global, ainda assim, não permite ao algoritmo retirar informação contida em cada uma das variáveis, sendo muito afectado pela imensa disparidade da variabilidade e posição relativa das distribuições não centradas[10];

c) p = 3 ou 4 no caso P2 (centragem por variável) – ao realizar esta operação agora todas as variáveis estão centradas e o algoritmo torna-se sensível à variabilidade das distribuições mais dispersas;

d) p = 4 (podendo também ser estendida a 7 ou 11) no caso P3 (normalização convencional) – neste caso as variáveis todas encontram-se em idêntico pé de igualdade e toda a informação útil e relevante pode ser descrita à custa do aumento do número de componentes principais a utilizar;

e) p = 3, 8 ou 12 no caso P5 (dados filtrados e PCA com normalização por variável) – a mesma observação pode ser extraída, contudo, o facto de os valores discrepantes terem sido removidos (ausência de outliers) faz com que as estimativas de dispersão e posição estejam agora correctas e as distribuições normalizadas apresentem características mais próximas da normalidade, requerendo mais componentes para descrever este sistema.

4.2.2 Impacto das variáveis

Nos casos P0 (sem qualquer tratamento) e P4 (normalização robusta) está-se reduzido à unidimensionalidade (p = 1), estando a informação útil excessivamente comprimida de modo que não permite a correcta racionalização do sistema em estudo. Nestes casos, as variáveis que apresentam maior impacto são:

a) em P0 (sem transformação) as variáveis V03 (P), V02 (Ba), V16 (Mn) e V05 (Cr), o que não deixa de ser curioso já que estas coincidem com as variáveis que possuem maior estimador central (543.65, 385.48, 283.91, 240.12) e estão, portanto, a condicionar os resultados devido ao seu efeito de escala;

b) em P4 (com normalização robusta) a variável V10 (Sb-Color.), por a estimativa robusta para a dispersão se aproximar de zero, provocando uma distribuição deficientemente normalizada, perdendo-se a

[10] Sem que as suas médias sejam nulas.

restante relevância da informação contida nas restantes variáveis.

Na tabela 4.5 apresentam-se listados os valores dos pesos, "Loads", que traduzem o impacto de cada variável original sobre a componente principal.

Tabela 4.5 - Impacto das variáveis sob a primeira componente principal (PC1) para os tratamentos P0 (sem qualquer tratamento) e P4 (com uma normalização robusta. Os loads mais significativos estão assinalados a negrito.

Var.	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13
P0	0.005	**0.483**	**0.710**	0.038	**0.295**	0.000	0.034	0.089	0.014	0.007	0.027	0.007	0.001
P4	0.000	0.000	0.000	0.000	0.000	0.009	0.003	0.000	0.109	**0.977**	0.024	0.006	0.000

Var.	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26
P0	0.042	0.132	**0.379**	0.002	0.002	0.020	0.007	0.002	0.010	0.010	0.001	0.006	0.001
P4	0.001	0.000	0.002	0.004	0.000	0.007	0.073	0.167	0.001	-0.001	0.000	0.000	0.001

Da tabela 4.5 verifica-se que, no caso P0 (sem tratamento), as variáveis mais relevantes para descrever o sistema são V03 [0.710], V02 [0.483], V16 [0.379] e V05 [0.295]. Esta mesma sequência está relacionada com a estimativa posicional (ordem de grandeza) da cada uma destas variáveis (543.65, 385.48, 283.91, 240.12), como pode ser visto da figura 4.4.

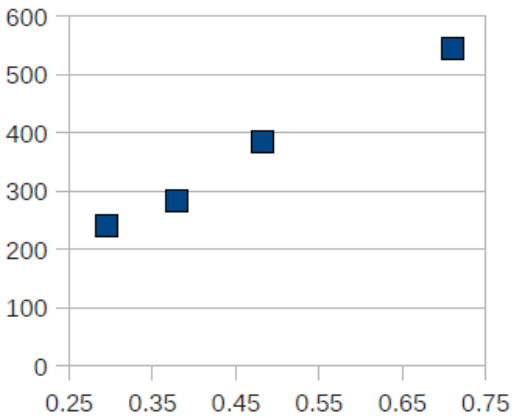


Figura 4.4 – Representação do valor médio das variáveis em relação ao respectivo peso (relevância) dessa variável sobre a primeira componente principal, no caso do processamento P0 (sem qualquer pré-tratamento).

De igual modo, para o caso do processamento P4 (com normalização robusta), apenas a variável V10 (Sb-Color.) se revela como mais significativa, corroborando o que foi anteriormente dito sobre a estimativa robusta da dispersão desta distribuição.

Quando se realiza a centragem global (P1) o número de componentes principais varia entre 1 e 2. Considerando estas duas primeiras componentes (PC1 e PC2), tabela 4.5, as variáveis com maior impacto são:

PC1 [91.23%] - V03 (P), V02 (Ba), V16 (Mn) e V05 (Cr) são as variáveis que possuem maior ordem de grandeza, como em P0:

PC2 [4.14%] - as variáveis V16 (Mn) e V02 (Ba).

Tabela 4.5 – Impacto das variáveis para o caso de pré-tratamento através de uma centragem global (caso P1).

Var.	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13
PC1	-0.091	**0.445**	**0.703**	-0.053	**0.233**	-0.096	-0.058	0.003	-0.081	-0.089	-0.066	-0.088	-0.096
PC2	0.053	**-0.423**	0.011	0.023	-0.043	0.057	0.022	0.029	0.055	0.069	0.056	0.054	0.057

Var.	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26
PC1	-0.049	0.052	**0.329**	-0.094	-0.095	-0.074	-0.088	-0.094	-0.086	-0.085	-0.096	-0.090	-0.095
PC2	0.027	-0.092	**0.867**	0.056	0.057	0.053	0.058	0.059	0.063	0.056	0.057	0.053	0.057

Da tabela 4.5 verifica-se que, quando se realiza a centragem global (P1), as variáveis que apresentam maior impacto - V03 (P, [0.703]), V02 (Ba [0.445, -0.423]), V16 (Mn, [0.329, 0.867]), V05 (Cr, [0.233]) - correspondem àquelas que também possuem médias maiores. Na figura 4.5 encontra-se representada a magnitude da diferença entre a média de cada variável em relação à média global (69.949).

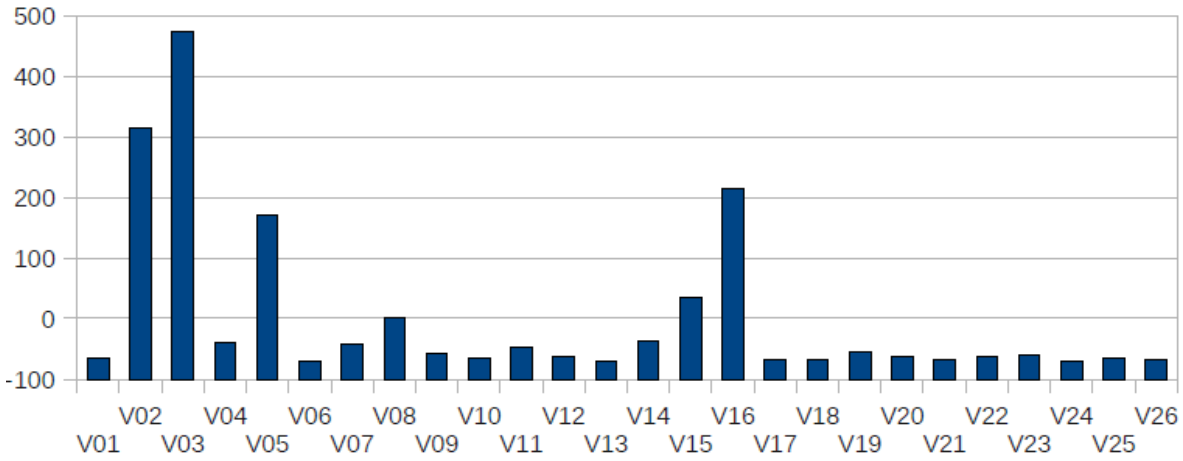


Figura 4.5 – Representação das diferenças da média de cada variável em relação à média global (69.949).

Da figura 4.5 verifica-se que as maiores diferenças entre médias (local e global) são obtidas nos casos V03 (473.7), V02 (315.5), V16 (213.9) e V05 (170.2). Se se representar de igual modo a amplitude da diferença de médias (local – global) em função do respectivo Load, como na figura 4.4, obtêm-se também uma dependência linear crescente.

No caso em que a centragem é feita variável a variável (P2), o número de componentes principais oscila entre 3 e 4. O impacto das variáveis encontra-se resumido na tabela 4.6.

Tabela 4.6 - Impacto das variáveis sobre as primeiras quatro componentes principais quando se efectua a centragem variável a variável (caso P2).

Var.	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13
PC1	0.001	0.075	**0.729**	0.015	-0.015	0.000	0.001	0.031	0.001	0.003	0.010	0.001	0.000
PC2	-0.003	**-0.384**	**-0.571**	-0.043	**0.258**	0.000	-0.023	-0.006	0.006	0.011	0.008	0.003	0.000
PC3	0.005	**0.619**	**-0.371**	0.048	**-0.572**	0.000	0.028	0.075	0.010	0.026	-0.003	-0.004	0.000
PC4	0.000	**0.645**	-0.071	0.013	**0.758**	0.000	0.018	-0.053	0.002	0.003	-0.004	0.003	0.000

Var.	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26
PC1	0.012	0.008	**0.679**	0.001	0.000	0.010	0.002	0.000	0.014	-0.001	0.000	0.000	0.001
PC2	-0.024	-0.127	**0.664**	-0.001	0.002	-0.002	0.004	0.005	-0.003	0.011	0.000	0.000	0.000
PC3	0.037	**0.215**	**0.309**	0.001	-0.004	-0.008	0.004	-0.001	0.010	-0.002	0.000	0.000	-0.001
PC4	0.014	0.015	0.023	-0.001	0.005	0.022	0.002	0.005	-0.003	-0.001	0.000	0.000	0.000

Sobre a primeira componente principal (PC1, VarX%[11] = 54.2 %) apresentam-se as contribuições mais significativas de V03 (P) e V16 (Mn) – estas são as variáveis com maiores dispersões (tabela 4.2). Já sobre a segunda componente (PC2, VarX% = 24.0 %) manifestam-se as variáveis V16 (Mn), V03 (P), V02 (Ba) e V05 (Cr), sendo estas duas últimas as variáveis que apresentam também grande variabilidade (tabela 4.2). Na terceira componente principal (PC3, VarX% = 12.3 %) encontram-se expressas as mesmas variáveis de PC2 acrescidas da contribuição de V15 (V) já que a sua dispersão também é significativamente alta. Sobre a quarta componente (PC4, VarX% = 4.5 %) apenas se destacam as contribuições de V16 (Mn) e V15 (V).

Neste caso, ao fazer a centragem variável a variável, a informação relevante encontra-se destacada sob a forma de variabilidade de cada distribuição, sendo as de maior variabilidade as que vão condicionar mais a análise PCA, ficando estas evidenciadas através do respectivo peso (Load) significativo que irá surgir nas componentes principais que descrevem o sistema.

Este processamento pode ser recomendado quando se pretende, por exemplo, evidenciar quais as variáveis que apresentam maior variabilidade. Caso se pretenda evitar esta perturbação é desejável realizar a normalização de cada variável.

Realizando a normalização de cada variável (P3 e P5) a situação torna-se mais difícil de ser interpretada já que a dimensionalidade do problema continua a exceder as limitações Humanas devido ao número de componentes principais poder ser superior a três.

Considerando o processamento P3 (normalização sem remoção prévia de outliers) a situação mais simples (p = 4) conduz a uma recuperação da variabilidade de apenas 53.8 %, o que é sobejamente insuficiente para descrever o sistema em causa. Se se assumir p = 7 a situação também não melhora muito (67.3 %) sendo necessário ir até p = 11 para se obter 80.9 % da reprodução dos resultados.

Na tabela 4.7 encontram-se resumidos o impacto de cada variável sobre as primeiras 7 componentes principais.

[11]VarX% representa a variância justificada por esta componente principal.

Tabela 4.7- Impacto das variáveis sobre as primeiras 7 componentes principais no caso em que a normalização paramétrica é realizada sobre cada variável (caso P3).

Var.	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13
PC1	**0.378**	**0.328**	0.164	**0.361**	**-0.293**	0.056	0.178	0.155	-0.009	-0.005	-0.011	-0.124	-0.030
PC2	0.008	0.001	0.030	0.007	0.013	0.128	0.006	0.065	**0.461**	**0.461**	**0.339**	0.115	-0.001
PC3	-0.029	-0.090	**0.378**	0.061	0.154	0.019	**-0.229**	**0.307**	-0.125	-0.127	0.144	0.053	0.050
PC4	0.037	-0.071	**-0.294**	0.016	-0.109	-0.089	**-0.364**	**0.421**	-0.045	-0.053	**0.202**	-0.175	-0.114
PC5	0.014	0.011	-0.157	-0.019	-0.071	0.130	0.135	0.186	-0.043	-0.039	-0.055	0.008	-0.095
PC6	0.025	0.039	-0.090	0.192	**0.247**	0.129	0.009	0.202	-0.016	-0.017	0.004	0.084	**0.525**
PC7	-0.008	-0.089	0.055	-0.038	-0.004	**0.316**	-0.004	0.024	-0.052	-0.052	0.106	**0.291**	**0.396**

Var.	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26
PC1	**0.319**	**0.387**	0.012	**0.211**	**-0.211**	0.008	-0.009	-0.056	**0.215**	-0.189	0.038	-0.054	-0.031
PC2	0.053	0.002	0.074	0.033	-0.018	0.064	**0.452**	**0.450**	0.015	0.024	0.034	-0.005	0.051
PC3	0.082	-0.125	**0.489**	0.008	0.138	0.194	-0.044	-0.076	**0.365**	0.192	**0.208**	0.053	**0.270**
PC4	0.115	0.113	-0.116	**-0.269**	**-0.287**	**-0.236**	-0.020	0.029	-0.142	**0.360**	**0.244**	-0.135	-0.099
PC5	**-0.243**	-0.039	**-0.225**	**0.369**	-0.058	**0.425**	0.018	-0.037	**-0.297**	0.004	**0.485**	0.067	**0.350**
PC6	0.151	0.045	-0.175	**-0.209**	**0.332**	0.083	0.001	0.000	-0.080	**-0.202**	**0.307**	**0.243**	**-0.368**
PC7	0.087	-0.027	-0.121	0.009	-0.079	-0.083	-0.090	-0.056	-0.083	-0.081	-0.080	**-0.711**	**0.236**

Das tabelas 4.4 e 4.7, através da análise detalhada dos Loads, verifica-se que a primeira componente (PC1, 21.3 %) está mais relacionada com as variáveis V15 (V), V01 (Fe), V02 (Ba), V14 (Ni), V05 (Cr), V22 (Co), V17 (Be) e V18 (Mo), podendo ser estas aquelas que possuem maior poder discriminante. Salienta-se, ainda, que as variáveis V17 (Be), V18 (Mo) e V22 (Co) contém casos severos de outliers, razão pela qual estas sejam talvez aqui distinguidas.

Na segunda componente principal (PC2, 16.5 %) encontram-se representadas as variáveis V09 (Sb), V10 (Sb-Color.), V20 (W), V21 (W-Color.) e V11 (Pb). Neste caso, à excepção das variáveis V21 e V11, aqui são evidenciadas variáveis sem grande informação (V09 e V20) e uma variável com casos severos de outliers (V10).

Sobre a terceira componente principal (PC3, 8.7 %) surgem as contribuições V16 (Mn), V03 (P), V08 (Zn), V26 (U), V07 (B), V22 (Co, com outliers) e V24 (Cd, sem informação). Já a componente seguinte (PC4, 7.4 %) repete as variáveis V03, V07, V08 de PC3 e acrescenta V11, V17, V18, V19, V23, V24. As restantes componentes têm contribuições inferiores e muito parecidas [4.84, 4.45, 4.15, 3.83, 3.44, 3.29 e 3.14%], próximas da contribuição basal [2.91, 2.75, 2.28, 1.88, 1.83,...], não sendo, por isso, consideradas.

Ordenando agora as variáveis mais representadas nas componentes principais, por ordem decrescente de contribuição na descrição da variabilidade do sistema, chega-se às seguintes constatações:

a) V15 (V), V01 (Fe), V02 (Ba), V14 (Ni), V05 (Cr), V22 (Co, com outliers), V17 (Be, com outliers), V18 (Mo, com outliers);

b) V09 (Sb, sem informação), V10 (Sb-Color., com outliers), V20 (W, sem informação), V21 (W-Color.) e V11 (Pb);

c) V16 (Mn), V03 (P), V08 (Zn), V26 (U), V07 (B) e V24 (Cd, sem informação);

d) V19 (As, com outliers), V23 (Y, com outliers).

Estes resultados não são fáceis de racionalizar por si só já que o algoritmo PCA está a descrever o sistema sem olhar à posição nem à variabilidade da distribuição. Nestas condições, ele anda a explorar a variabilidade intrínseca de cada variável, algo que nos é difícil de avaliar de outro modo. Contudo, através da matriz de correlação destas variáveis, tabela 4.7, pode-se tentar perceber melhor o que se está a passar.

Tabela 4.8 – Matriz de correlação das variáveis (diagonal inferior) referentes ao caso em que a normalização paramétrica é realizada sobre cada variável (caso P3). Valores mais relevantes assinalados a negrito.

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V14	**V15**	V16	V17	V18	V19	V20	V21	V22	V23	V24
V02	**0.66**																						
V03	0.24	0.26																					
V04	**0.73**	0.64	0.37																				
V05	**-0.56**	-0.49	-0.09	-0.47																			
V06	0.08	0.03	0.07	0.07	-0.02																		
V07	0.31	0.37	0.10	0.36	-0.24	0.06																	
V08	**0.41**	0.14	0.11	0.30	-0.21	0.04	-0.18																
V09	0.03	0.03	0.01	0.02	-0.06	-0.02	0.14	0.00															
V10	0.20	0.18	0.08	0.18	-0.24	0.05	0.47	-0.17	0.23														
V11	-0.07	-0.09	0.12	-0.06	0.08	0.12	-0.26	**0.49**	0.06	-0.10													
V12	-0.20	-0.18	0.01	-0.19	0.26	-0.03	-0.01	-0.11	0.03	-0.01	-0.02												
V14	**0.69**	**0.47**	0.28	**0.68**	-0.32	0.17	0.15	0.41	0.00	0.12	0.11	-0.14											
V15	**0.87**	**0.74**	0.15	**0.76**	**-0.64**	0.06	0.37	**0.40**	0.05	0.20	-0.01	-0.25	**0.74**										
V16	-0.04	-0.05	**0.41**	-0.02	0.08	0.00	-0.08	0.08	-0.03	-0.05	0.16	0.10	0.09	-0.13									
V17	**0.38**	0.35	0.29	0.32	-0.35	0.10	0.41	0.02	0.08	0.29	-0.07	-0.09	0.21	0.35	0.08								
V18	**-0.42**	-0.29	0.02	-0.28	**0.54**	-0.04	-0.05	-0.26	-0.03	-0.02	-0.04	0.20	-0.27	**-0.45**	0.08	-0.13							
V19	0.08	0.06	0.18	0.00	0.08	0.08	-0.02	0.00	-0.01	0.02	0.06	0.06	-0.05	-0.05	0.09	0.20	0.08						
V20	0.00	-0.02	0.06	-0.01	0.02	0.12	-0.07	0.10	-0.02	-0.01	0.12	0.04	0.01	0.01	0.17	0.06	0.03	0.15					
V21	-0.22	-0.23	-0.12	-0.23	0.35	-0.08	-0.18	0.14	0.05	-0.19	0.18	0.14	-0.15	-0.20	0.04	-0.24	0.00	0.05	0.14				
V22	**0.38**	0.28	**0.51**	**0.48**	-0.24	0.05	0.10	0.14	-0.04	0.04	0.01	-0.10	**0.47**	0.31	**0.43**	0.19	-0.08	0.07	0.11	-0.18			
V23	**-0.37**	-0.34	-0.26	-0.38	0.25	-0.14	**-0.43**	0.14	-0.03	-0.20	0.27	0.05	-0.20	-0.32	0.22	-0.23	0.10	-0.07	0.14	0.29	-0.20		
V24	0.05	0.07	0.03	0.02	-0.09	0.09	-0.04	0.06	-0.02	0.03	0.08	0.02	0.01	0.02	0.09	0.05	-0.05	0.13	0.24	0.02	0.16	0.08	
V26	-0.06	-0.09	0.18	-0.11	0.11	0.08	-0.07	0.07	0.04	0.02	0.12	0.13	-0.12	-0.17	0.08	0.11	0.06	0.14	0.17	0.04	-0.02	0.14	0.12

Da tabela 4.8 verifica-se que existem algumas variáveis que apresentam correlações significativas[12] com outras variáveis. São exemplos as variáveis V15 (V, com 7 correlações significativas), a V14 e a V22 (Ni e W-Color., com 5 correlações significativas) e a V18 (Mo, com 3 correlações significativas).

Se atentarmos, por exemplo, na variável com maior número de correlações significativas, a variável V15 (V), tabela 4.8, verifica-se que esta está muito correlacionada com V01 (Fe), V02 (Ba), V04 (Cu), V05

[12]Através de um teste t-student unilateral, verifica-se que, com n = 516 valores e ao nível de confiança de 95 %, uma correlação torna-se significativa quando o seu valor excede, em módulo 0.40.

(Cr), V14 (Ni), V17 (Be, com outliers), V18 (Mo, com outliers), V22 (Co, com outliers). Comparando agora com a informação contida na primeira componente principal (PC1), tabela 4.7, verifica-se que sobre esta componente estão representadas as contribuições maioritárias de V01, V02, V04, V05, V14, **V15**, V17, V18 e V22 – há uma enorme coincidência entre as correlações observadas e as contribuições aparentes sobre cada componente. Saliente-se ainda que o sinal do peso (Load) está directamente relacionado com o tipo de correlação verificada – Load negativo para o caso de correlações negativas. Relações similares podem também ser observadas para as restantes componentes.

Esta constatação faz-nos reflectir sobre o algoritmo PCA. Sendo a matriz de correlação a base de todo o processamento da decomposição singular sobre as variáveis do sistema, parece que as componentes principais resultam da análise da variabilidade intrínseca das variáveis do sistema (normalizadas). Assim, o algoritmo procura descrever o máximo de informação com o mínimo de componentes principais, colocando sobre cada uma delas parte da informação mais correlacionada entre si. As componentes seguintes vão descrevendo, de modo similar, com menor impacto, as restantes correlações, de forma a decompor a informação de cada variável em contribuições lineares interdependentes até esgotar a informação relevante, sob a forma de correlação, de todas as variáveis. Assim, o PCA tem tendência a dar menos importância a variáveis aleatórias independentes, concentrado a sua atenção nas variáveis mais correlacionadas.

De acordo com estas observações, a análise PCA revela-se importante para, numa análise multivariada a um sistema desconhecido, avançar interdependências de variáveis, podendo eventualmente ser utilizada como diagnóstico de factores.

Após estas conclusões importa ainda verificar se a presença de outliers é ou não prejudicial. Ou seja, pretende-se averiguar o facto de que, sendo estes valores discrepantes capazes de terem um forte impacto na estimativa da variabilidade de algumas distribuições, até que ponto a análise PCA vem prejudicada ou afectada por este enviesamento de estimativas.

Assim, realizou-se uma nova abordagem PCA, após a remoção dos valores discrepantes identificados.

Anteriormente já se verificou que, no caso da análise com P5 (após remoção de 15 outliers), o número de componentes principais pode ser considerado p = 3, 8 ou 12, dependendo do critério assumido. Na tabela 4.9 encontra-se um resumo do impacto das variáveis sobre cada componente principal.

Tabela 4.9- Impacto das variáveis a partir da normalização de cada variável, na ausência de valores discrepantes (caso P5).

Var.	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12
PC1	**0.369**	**0.324**	0.159	**0.356**	**-0.291**	0.056	0.196	0.135	0.028	0.132	-0.027	-0.116
PC2	0.046	-0.044	**0.220**	0.039	0.068	0.073	-0.313	**0.402**	-0.065	-0.230	**0.405**	0.016
PC3	-0.092	-0.011	**0.389**	-0.007	0.143	0.100	0.219	**-0.280**	0.050	0.222	-0.112	0.203
PC4	0.035	0.011	-0.230	-0.123	-0.138	0.185	0.146	0.091	**0.327**	**0.312**	0.175	0.045
PC5	0.002	-0.035	0.127	0.068	0.174	-0.222	0.173	0.164	**0.536**	0.237	0.237	0.243
PC6	0.087	0.012	0.031	0.088	**0.314**	**0.518**	0.014	0.129	-0.183	-0.165	0.145	0.142
PC7	0.137	0.175	-0.119	0.118	0.147	**-0.387**	0.103	-0.044	-0.108	-0.123	**-0.324**	**0.416**
PC8	0.075	0.088	0.175	-0.048	-0.105	**-0.451**	-0.138	0.090	-0.088	-0.218	-0.014	-0.097

Var.	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25
PC1	**0.308**	**0.379**	-0.001	**0.213**	**-0.201**	0.012	-0.005	-0.149	**0.210**	**-0.203**	0.025	-0.049
PC2	0.177	0.016	**0.331**	-0.055	-0.046	0.122	0.256	0.202	0.229	0.272	0.177	0.174
PC3	-0.088	-0.185	**0.317**	0.271	**0.303**	**0.278**	0.147	-0.126	0.251	-0.158	0.114	**0.232**
PC4	-0.141	0.049	-0.241	0.228	-0.179	0.191	0.288	0.158	**-0.334**	0.121	**0.276**	**0.316**
PC5	0.140	0.035	0.107	-0.028	0.116	-0.173	-0.251	0.260	-0.040	0.046	**-0.414**	-0.026
PC6	0.155	0.033	**-0.349**	-0.060	0.216	**0.285**	-0.097	0.084	-0.187	**-0.333**	-0.232	0.059
PC7	0.045	0.128	-0.069	-0.129	0.031	0.228	0.248	**0.469**	0.000	-0.026	0.192	-0.151
PC8	-0.214	-0.037	-0.075	0.267	-0.150	**0.438**	-0.281	-0.033	-0.154	-0.011	**-0.285**	**0.350**

Analisando em detalhe as contribuições das variáveis sobre as componentes principais, tabela 4.9, verifica-se que sobre a primeira componente (PC1, 23.8 %) as contribuições que mais se destacam são, por ordem decrescente de relevância V15 (V), V01 (Fe), V04 (Cu), V02 (Ba), V14 (Ni), V05 (Cr), V17 (Be), V22 (Co), V23 (Y) e V18 (Mo). Da segunda componente (PC2, 10.3%) destacam-se as contribuições de V11 (Pb), V08 (Zn), V16 (Mn), V07 (B), V14 (Ni), V03 (P), V17 (Be), repetindo ainda a V22 (Co). Já a terceira componente (PC3, 8.6 %) começa a repetir algumas das variáveis anteriores, introduzindo as novas contribuições de V18 (Mo), V19 (As), V25 (Nb) e V10 (Sb-Color.).

Com base apenas nestas três componentes principais consegue-se recuperar apenas 42.7% da variabilidade total o que ainda é pouco para descrever o sistema em causa mas, parece que a informação pertinente já está contida nestas variáveis já que, através do "scree plot", figura 4.3, a quarta componente já se confunde com a contribuição basal, com sequência de contribuições 6.11, 5.06, 4.67, 4.37, 4.17 %, respectivamente. Desta observação resulta que a série de variáveis identificadas nestas três primeiras componentes deve ser as que possuem mais informação sobre os dados e por isso as mais susceptíveis de serem utilizadas na avaliação e comparação destas amostras.

De igual modo, para que se possa fazer o mesmo tipo de verificação que no caso anterior, apresenta-se na tabela 4.10 os respectivos valores de correlação das variáveis para esta abordagem.

Tabela 4.10- Matriz de correlação das variáveis (diagonal inferior) referentes ao caso em que a normalização paramétrica é realizada sobre cada variável em caso de não estarem presentes valores discrepantes (caso P5). Valores mais relevantes assinalados a negrito.

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V14	**V15**	V16	V17	V18	V19	V20	V21	V22	V23	V24
V02	**0.68**																						
V03	0.24	0.26																					
V04	**0.75**	0.64	0.37																				
V05	**-0.60**	-0.49	-0.09	-0.47																			
V06	0.08	0.03	0.07	0.07	-0.02																		
V07	0.31	0.37	0.10	0.36	-0.24	0.06																	
V08	**0.43**	0.14	0.11	0.30	-0.21	0.04	-0.18																
V09	0.03	0.03	0.01	0.02	-0.06	-0.02	0.14	0.00															
V10	0.20	0.18	0.08	0.18	-0.24	0.05	0.47	-0.17	0.23														
V11	-0.07	-0.09	0.12	-0.06	0.08	0.12	-0.26	**0.51**	0.06	-0.10													
V12	-0.20	-0.18	0.01	-0.19	0.26	-0.03	-0.01	-0.11	0.03	-0.01	-0.02												
V14	**0.73**	**0.51**	0.31	**0.77**	**-0.41**	0.21	0.15	**0.38**	0.00	0.12	0.11	-0.14											
V15	**0.88**	**0.75**	0.15	**0.77**	**-0.65**	0.06	0.33	**0.43**	0.05	0.22	-0.01	-0.28	**0.77**										
V16	-0.04	-0.05	**0.41**	-0.02	0.08	0.00	-0.08	0.08	-0.03	-0.05	0.16	0.10	0.09	-0.13									
V17	**0.37**	0.35	0.29	0.32	-0.35	0.10	0.41	0.02	0.08	0.29	-0.07	-0.09	0.20	**0.35**	0.08								
V18	**-0.44**	-0.30	0.02	-0.28	**0.54**	-0.04	-0.05	-0.26	-0.03	-0.02	-0.04	0.20	-0.27	**-0.46**	0.08	-0.13							
V19	0.08	0.06	0.18	0.00	0.08	0.08	-0.02	0.00	-0.01	0.02	0.06	0.06	-0.05	-0.05	0.09	0.20	0.08						
V20	0.00	-0.02	0.06	-0.01	0.02	0.12	-0.07	0.10	-0.02	-0.01	0.12	0.04	0.01	0.01	0.17	0.06	0.03	0.15					
V21	-0.22	-0.23	-0.12	-0.23	0.35	-0.08	-0.18	0.14	0.05	-0.19	0.18	0.14	-0.15	-0.20	0.04	-0.24	0.00	0.05	0.14				
V22	**0.38**	0.28	**0.53**	**0.51**	-0.24	0.05	0.10	0.14	-0.04	0.04	0.01	-0.10	**0.49**	**0.38**	**0.44**	0.19	-0.08	0.07	0.11	-0.18			
V23	**-0.36**	-0.32	-0.26	**-0.38**	0.25	-0.14	**-0.43**	0.14	-0.03	-0.20	0.27	0.05	-0.20	**-0.39**	0.22	-0.23	0.10	-0.07	0.14	0.29	-0.20		
V24	0.05	0.07	0.03	0.02	-0.09	0.09	-0.04	0.06	-0.02	0.03	0.08	0.02	0.01	0.02	0.09	0.05	-0.05	0.13	0.24	0.02	0.16	0.08	
V26	-0.06	-0.09	0.18	-0.11	0.11	0.08	-0.07	0.07	0.04	0.02	0.12	0.13	-0.12	-0.17	0.08	0.11	0.06	0.14	0.17	0.04	-0.02	0.14	0.12

Como primeira análise da tabela 4.10 verifica-se que as estimativas de correlação não foram muito afectadas pela contaminação de 2.9 % de dados (15 outliers em 516 objectos), podendo ser identificados os mesmos padrões de correlação significativa.

Se se atentar de novo na tabela 4.9, verifica-se que sobre a primeira componente estão representadas as variáveis V01, V02, V04, V05, V14, **V15**, V17, V18, V22 e V23 o que está uma vez mais de acordo com as correlações significativas encontradas para a variável V15.

4.2.3 Representação dos objectos

Sendo o algoritmo PCA adequado para permitir a redução da dimensionalidade de um sistema multivariado, faculta, deste modo, a interpretação do sistema permitindo representar parcialmente a

informação contida no sistema num número muito restrito de dimensões – componentes principais. Este tipo de gráfico designa-se de "gráfico de scores" já que representa-se nesse sub-espaço os objectos em causa.

Assim, importa agora tentar visualizar a distribuição espacial das amostras no sub-espaço das respectivas componentes principais e, deste modo, ter uma noção do arranjo e proximidades de objectos, no sentido de identificar grupos mais densos, característicos de amostras mais homogéneas.

Na figura 4.6 estão representados gráficos de "scores" dos objectos considerando apenas as duas primeiras dimensões (2D).

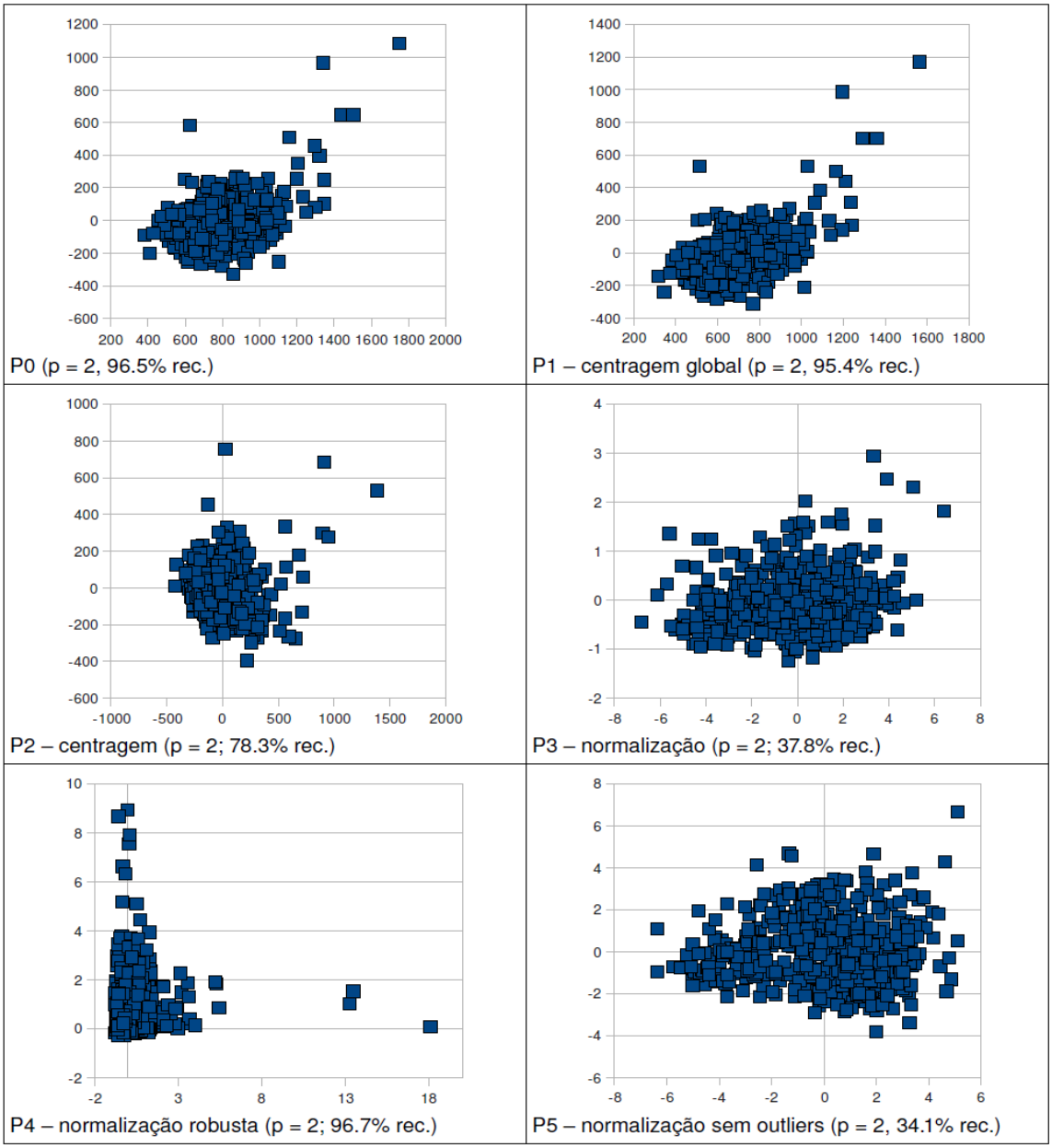


Figura 4.6 – Efeito do tipo de pré-processamento na representação bidimensional dos objectos no sub-espaço das componentes principais.

Tendo a perfeita noção de que, em alguns casos (P3 e P5), o número de componentes necessárias à representação do sistema é maior que dois, ainda assim, pode-se verificar da figura 4.6 que o sistema em

análise aparenta ser denso e coeso, observando-se diversos objectos sobrepostos sobre a mesma localização gráfica. Esta observação é, em parte, coerente com o facto se trata de um conjunto de amostras de sedimentos recolhidas na mesma região do país, sujeitas ao mesmo tipo de acção meteorológica, geofísica e ambiental, não permitindo tirar, deste modo, grandes conclusões sobre eventuais grupos ou famílias de amostras.

Desta figura observa-se ainda que, nos casos em que o sistema está sobejamente bem representado (P0, P1 e P2), o aspecto da distribuição das amostras é denso, com distribuição aleatória em torno de um valor médio único. Nota-se ainda que existem um conjunto de amostras que se afastam deste núcleo denso podendo ser revelação de anomalias ambientais ou outras causas a averiguar.

Observa-se, ainda, que no caso do processamento com estimativas robustas (P4), a informação está distorcida devido à anomalia anteriormente identificada, relacionada com uma estimativa muito pequena para a variabilidade dos dados.

No sentido de perceber ainda mais acerca deste sistema encontra-se representada a mesma informação sobre o aspecto tridimensional, figura 4.7.

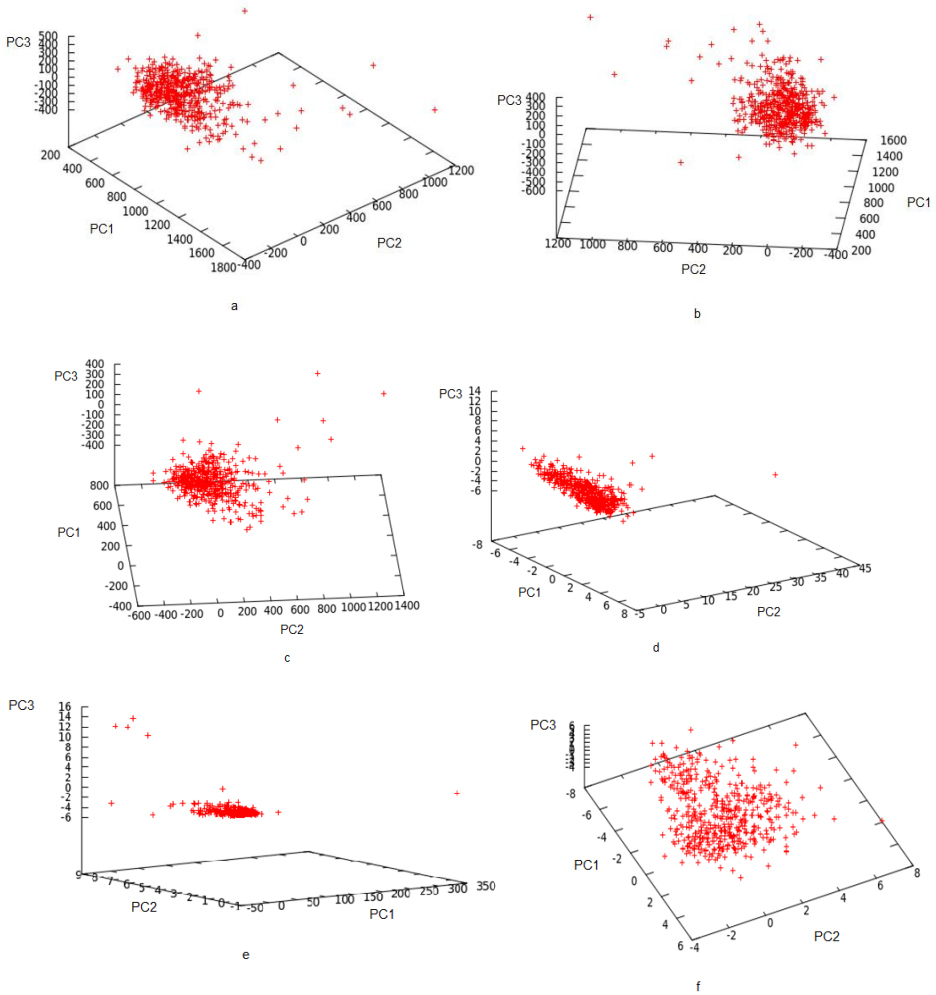


Figura 4.7 – Efeito do tipo de pré-processamento na representação bidimensional dos objectos no sub-espaço das componentes principais: pré-processamentos a) P0, b) P1, c) P2, d) P3, e) P4 e f) P5.

A figura 4.7 permite ainda confirmar que, de facto, as amostras parecem ser muito similares em composição havendo alguns casos esporádicos que devem ser analisados em maior detalhe, no sentido de permitirem a identificação de contaminações ou outras causas físico-químicas.

A nossa atenção vai ainda para os pré-processamentos que envolvem a normalização, P3 (n = 516, amostra original) e P5 (n = 501, após a remoção de 15 outliers identificados) já que no último caso, a remoção dos valores discrepantes, alguns bem patentes na figura (d), permite a expansão da escala e melhor visualização espacial dos objectos, verificando-se que, ao rodar os eixos parece existir uma estrutura similar a três sub-grupos que poderão ser averiguados no sentido de reconhecer proveniências e eventuais exposições às condições externas.

Saliente-se ainda que, no caso P5, embora a representação tridimensional em causa apenas traduza 42.8 % da informação original, esta representação parece estar suficientemente bem representada já que o mesmo padrão, conjunto de objectos dispersos de forma mais ou menos aleatória, é coerente com as restantes abordagens.

Esta última figura revela ainda que, a representação dos dados, ainda que de uma forma insuficiente, possibilita alguma percepção sobre o sistema multivariado em análise, que de outro modo seria inconcebível representar, facultando alguma racionalização do sistema em análise.

Em jeito de resumo e conclusão, a abordagem PCA desenvolvida (P0 a P5) permitiu-nos extrair informação acerca do efeito do pré-processamento na extracção de informação útil do sistema a diferentes graus – em termos de posição, em termos de dispersão e ainda em termos de correlação (interdependência das variáveis).

A representação dos objectos através da análise PCA foi devidamente esclarecedora já que permitiu visualizar eventuais objectos discrepantes e verificar o efeito do pré-tratamento na descriminação da informação – mais discriminante no caso da normalização correcta, sem efeito de outliers.

4.2.4 Análise de interdependências

Da análise PCA resultou que foram detectadas algumas inter-relações entre variáveis, estatisticamente significativas. Aqui algumas explicações para este fenómeno podem ser avançadas. A primeira faz-nos pensar que se tratando de sedimentos colhidos numa região (localidade) é natural que este confinamento espacial e físico-químico tenha grande impacto sobre a homogeneidade das amostras resultando uma interdependência nas variáveis analisadas, que se pretende que sejam discriminadoras. A segunda eventual explicação para este fenómeno pode estar relacionada com o facto de alguns dos elementos em análise (Cr, V, Mn, U, V, Fe, Al, Co, Cu, entre outros) terem tendência para atingir naturalmente, em ambiente oxidante, estados de oxidação mais elevados originando os respectivos óxidos

que se comportam em solução como aniões. Estes, por sua vez, têm tendência para precipitar por combinação com outros catiões mais simples (Ca, Mg, Ba, entre outros). Uma terceira eventual explicação para estas interdependências pode estar relacionada com o facto de que as técnicas utilizadas (ICP e colorimetria) são susceptíveis à interferência mútua com os elementos presentes na matriz da amostra.

Numa tentativa derradeira de averiguar se esta dependência tem alguma notoriedade, utilizou-se o algoritmo PLS para averiguar estas dependências.

Partiu-se da sub-amostra que foi tirada no sentido da remoção dos valores discrepantes. Todas as variáveis foram devidamente normalizadas com base nas suas estimativas paramétricas (média e desvio padrão) correspondendo ao pré-processamento P5 da análise PCA.

O sub-espaço dos predictores foi construído com todas as variáveis à excepção de V15 (V) que foi isolada como resposta única.

De modo similar à análise PCA, a análise PLS revela que basta apenas considerar os primeiros três factores latentes para conseguir descrever a maioria da informação contida na resposta V15 (V). O primeiro factor latente é de igual modo o mais expressivo já que é capaz de descrever 82.0 % da informação contida na resposta. Os restantes factores latentes (FL2 e FL3) são responsáveis pela descrição de 6.4 e 1.4 % da informação contida nessa resposta.

Saliente-se ainda que a descrição integral desta variável só é possível se se aumentar o número de factores latentes para 13.

Na tabela 4.11 apresenta-se um resumo do impacto dos predictores (Loads) sobre a resposta em causa.

Tabela 4.11- Impacto das variáveis na descrição da resposta V15 através do algoritmo PLS.

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11
	Fe	Ba	P	Cu	Cr	Ag	B	Zn	Sb	Sb-Color.	Pb
FL1	0.440	0.375	0.075	0.387	-0.327	0.033	0.187	0.202	0.024	0.103	-0.003
FL2	0.189	0.122	-0.494	-0.033	-0.013	-0.146	-0.103	0.215	-0.029	-0.198	0.117
FL3	0.243	0.269	-0.040	-0.073	0.092	-0.058	0.304	-0.351	-0.004	0.067	-0.093

	V14	V16	V17	V18	V19	V20	V21	V22	V23	V24	V26
	Ni	Mn	Be	Mo	As	W	W- Color.	Co	Y	Cd	U
FL1	0.374	-0.064	0.175	-0.231	-0.026	0.005	-0.104	0.160	-0.164	0.012	-0.088
FL2	0.186	-0.290	-0.278	-0.029	-0.169	0.084	0.302	-0.359	0.301	-0.075	-0.148
FL3	0.204	0.144	0.073	0.555	0.172	0.272	0.060	-0.066	0.099	0.003	0.123

Da tabela 4.11 verifica-se que no primeiro factor latente, o mais relevante já que consegue descrever 82.0 % da informação contida em V15, estão expressas as dependências com V01, V02, V04, V05, V14, V17, V18, V22 e V23. Este resultado está em conformidade com o que foi anteriormente observado acerca do tratamento PCA no caso P5, discutido a partir das tabelas 4.10 e 4.11.

Esta observação permite-nos agora afirmar que, cerca de 82.0 % da informação contida em V15 (V)

depende linearmente das restantes variáveis observadas.

De igual modo foi também feito um estudo para as variáveis V14 (Ni), V18 (Mo) e V22 (W-Color.).

No primeiro caso, V14, os primeiros quatro factores latentes conseguem descrever um total de 70.3 % da informação contida nesta variável sendo as respectivas contribuições de 53.8, 10.3, 4.4 e 1.9 %. Através do factor latente mais significativo, esta variável aparenta estar muito relacionada com V01 [0.430], V02 [0.296], V04 [0.425], V08 [0.256], V15 [0.458] e V22 [0.290]. Não deixa de ser interessante verificar aqui a coexistência das variáveis V01, V02, V04 e V22 também a descrever esta variável. Já a contribuição de V15 seria de esperar, uma vez que V14 é uma das que está patente na descrição de V15.

Já no caso da variável V18, esta não contem muita informação dependente, conforme visto anteriormente. São necessários três factores latentes para conseguir descrever cerca de 38.7 % da informação contida nesta variável sendo as contribuições individuais sucessivas 23.9, 10.3 e 4.5 %. O primeiro factor latente revela o impacto de V01 [0.401], V02 [0.283], V04 [0.273], V05 [-0.523], V14 [0.261] e V15 [0.436]. De novo surgem as variáveis V01, V02, V04, V05 na descrição de outras variáveis. As variáveis V14 e V15 estão aqui também relacionadas com a V18.

Passando por fim à variável V22, são também necessários três factores latentes para descrever cerca de 51.3 % da sua informação, sendo as respectivas contribuições de 34.3, 14.2 e 2.7 % nessa descrição. Olhando de novo às contribuições manifestadas através do primeiro factor latente, as variáveis mais significativas nesta descrição são V01 [0.289], V02 [0.233], V03 [0.422], V04 [0.397], V14 [0.387], V15 [0.261] e V16 [0.357]. Saliente-se, uma vez mais, a contribuição das variáveis V01, V02 e V04 nesta representação linear, também as variáveis V03 e V16 e as já conhecidas V14 e V15.

4.3 Variáveis relevantes

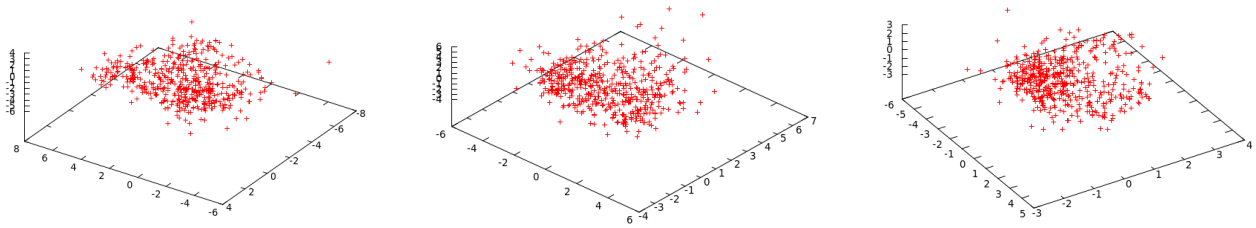
Da anterior discussão breve sobre este sistema de amostras de sedimentos, ficou-se com a noção de que, por alguma das razões anteriormente avançadas, existem variáveis que parecem ser fundamentais na descrição deste conjunto de dados, enquanto que outras são apenas dependentes destas primeiras, havendo ainda um conjunto de variáveis que não aparentam ser discriminantes (não contribuem para a distinção dos objectos).

Na tentativa de tentar verificar esta hipótese executou-se, uma vez mais, a análise PCA, como no caso P5, mas acrescentou-se ainda duas condicionantes. No primeiro caso (P5a) efectuou-se apenas a exclusão das variáveis mais dependentes (V14, V15, V18 e V22) do conjunto inicial, no caso seguinte (P5b) apenas foram consideradas as variáveis que se julgam ser mais relevantes (V01, V02, V03, V04 e V05).

Como resultados da análise PCA obteve-se que no primeiro caso (P5) com três componentes principais consegue-se descrever cerca de 42.8 % da informação do sistema. Ao remover alguma informação correlacionada (procedimento P5a) com o mesmo número de componentes consegue-se descrever cerca de 48.7 % dessa informação, já que também o número de variáveis (quantidade de

informação inicial) foi ligeiramente reduzida tendo-se passado de m = 24 para m = 20. Com a seguinte simplificação (P5b), apenas considerando algumas das variáveis menos correlacionadas passa-se para m = 5 e, neste caso a quantidade de informação descrita com as três primeiras componentes passa agora a ser de 87.5 %.

Na figura 4.8 encontram-se representadas os respectivos gráficos tridimensionais dos scores.



P5 (m = 26) P5a (m = 20) P5b (m = 5)

Figura 4.8 – Representação tridimensional dos objectos no sub-espaço das componentes principais.

Da figura 4.8 verifica-se que as representações são todas similares, traduzindo, essencialmente, a mesma noção de que as amostras apresentam uma distribuição densa e, mais ou menos, aleatoriamente distribuída em torno de um valor central.

Deste modo, fica-se com a noção de que é possível simplificar os problemas multivariados se se realizar uma análise criteriosa às independências das respectivas variáveis, procurando apenas reter aquelas que são independentes e apresentam maior poder discriminante. Assim, como tal, pode-se manter a informação pertinente do sistema e retirar o mesmo tipo de conclusões com menor dispêndio de esforço, o mais completo possível e de um modo mais económico, por exemplo, minimizando o número de variáveis (análises laboratoriais).

Capítulo 5

Conclusão

Este trabalho de projecto permitiu extrair algumas conclusões deveras importantes quando se utiliza os algoritmos PCA e PLS na análise de sistemas multivariados.

Dado que ambos algoritmos requerem, numa fase inicial de pré-acondicionamento, a utilização de estimativas paramétricas (média e desvio padrão), fez-se uma análise prévia de cada variável, no sentido de detectar valores discrepantes, que condicionam estas estimativas. Verificou-se também que a estimativa da dispersão era aquela que mais sofria inflação com a presença dos valores discrepantes.

Ainda quanto ao pré-processamento a realizar (nenhum, centragem ou normalização) verificou-se que estes vão condicionar de forma determinante a sensibilidade do algoritmo na análise multivariada – sem centragem a informação está de tal forma condensada que não é inteligível; a centragem global torna o PCA sensível às estimativas da média de cada distribuição; a centragem por variável faz com que este passe a dar relevância à variabilidade (amplitude de variação) de cada variável. Já a normalização, procedimento que se recomenda, coloca o algoritmo a explorar as verdadeiras inter-relações entre variáveis, indo à sub-estrutura destas.

O facto de haver outliers presentes, em determinadas variáveis, é um facto que vai condicionar as estimativas paramétricas e adulterar um pouco a normalização. Contudo, ficou demonstrado que utilizar estimadores robustos para o mesmo efeito pode ser prejudicial, já que também pode levar a deformações na estrutura (Loads) e, consequentemente, na representação dos dados (gráficos dos scores).

Deste estudo ficou ainda patente que, no sistema estudado, existem correlações significativas entre variáveis que podem ter pelo menos três causas (confinamento físico-químico, co-precipitação e/ou eventual interferência espectral). Assim, pode-se dizer que há dois tipos extremos de variáveis: as fundamentais, que transportam a informação relevante, e as variáveis dependentes, que são redundantes - nada acrescentam na descrição do sistema. No meio ficam as pseudo-variáveis, que apresentam baixa variabilidade, e as variáveis aleatórias, cuja contribuição não tem um sentido bem definido.

A matriz dos lambdas revela que existem 12 variáveis com expressão significativa, uma vez que estas estão fortemente correlacionadas, atendendo ao critério de Pearson. Por sua vez, a matriz de correlação revelou que existem variáveis que apresentam correlação significativa, do ponto de vista estatístico e geoquímico, entre si, sendo, nalguns casos, positiva (V com Fe, Cu, Ba e Ni; Fe com Cu, Ni, Ba e Zn; Ni com Cu; Cr e Mo; Zn e Pb) e noutros negativa (V e Cr; Fe e Cr). Assim, a maioria das associações verificadas é entre elementos que possuem o mesmo comportamento geoquímico ou que precipitam num ambiente superficial.

Na verdade, os iões Cd^{2+}, Zn^{2+}, Cu^{2+}, Co^{2+}, Ni^{2+}, Sc^{3+}, Fe^{3+}, V^{5+} e Mn^{4+} possuem potenciais iónicos (z/r) intermédios (o log z/r varia entre 0.4 e 1), sendo insolúveis nas águas doces naturais, pelo que se

concentram na matéria em suspensão, por adsorção, sendo, desta forma, facilmente incorporados nos sedimentos de corrente. Nas águas naturais podem formar-se vários complexos de iões metálicos, podendo estes iões ser classificados tendo em conta a formação de complexos nas soluções aquosas, em 3 grupos: iões de metais – A ou duros, iões metais - B ou macios e iões de metais de transição [90 e 91].

Os iões de metais - A possuem simetria esférica, com nuvens electrónicas que dificilmente são deformadas por campos eléctricos, podendo ser vistas como "esferas duras". A estabilidade dos seus complexos aumenta com o seu potencial iónico. Já os iões dos metais alcalinos formam complexos instáveis, enquanto que o Zn forma complexos estáveis, podendo encontrar-se, neste grupo, o Ba^{+}.

Os iões dos metais - B (Cu^{2+}, Ag^{+}, Au^{+}, Tl^{+}, Ga^{+}, Zn^{2+}, Cd^{2+}, Hg^{2+}, Pb^{2+}, Sn^{2+}, Au^{3+}, In^{3+}, Bi^{3+}) possuem nuvens electrónicas não esféricas, sendo facilmente deformadas por campos eléctricos de outros iões, sendo muitos dos seus complexos muito pouco solúveis ou insolúveis.

Por sua vez, os iões de metais de transição (V^{2+}, Cr^{2+}, Mn^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Ti^{3+}, V^{3+}, Cr^{+}, Mn^{3+}, Fe^{3+}, Co^{3+}) possuem nuvens electrónicas não esféricas, contudo, apesar de não serem tão facilmente deformadas como as dos metais - B, o comportamento dos seus complexos é semelhante.

Até agora, consegue-se explicar as relações encontradas, por exemplo a de V-Fe, em que o V^{2+} compete com Fe^{2+} na malha dos minerais, como é o caso dos silicatos, por outro lado, as ligações V-Ba e Fe-Ba resultam de mineralizações de sulfuretos de Fe, V, entre outros, pelo que o ião sulfato estará presente nas águas, levando à formação de sulfato de bário, que é um sal insolúvel, logo, quanto maior for a quantidade de sulfuretos, maior será a relação entre o Fe, o V e o Ba.

No caso dos seguintes elementos V-Cu, V-Ni, Fe-Cu, Fe-Ni, Fe-Zn, Ni-Cu, as relações verificadas podem ser explicadas pela formação de complexos insolúveis.

No entanto, a relação entre os elementos Fe-Cr e V-Cr, apresenta correlações negativas porque o Cr^{3+} pode substituir o Fe^{3+} na malha dos minerais, como acontece nos óxidos, por isso, caso tenhamos minerais silicatos e formação de óxidos, há competição entre eles, verificando-se a oxidação dos silicatos e a produção de óxidos.

Bibliografia

1. Relatório do Estado do Ambiente 1999 — Solos; APRH - Associação portuguesa dos recursos hídricos;
2. Poluição do solo: revisão generalista dos principais problemas; Rodrigues S., Duarte A. C., Castro A., Santos T. (Ed.); O Ambiente e a Saúde; Lisboa, Instituto Piaget, pp. 136-176; 2003;
3. Environmental impact of metals derived from mining activities: Processes, predictions, prevention; Salomons W.; Journal of Geochemict Exploration; 52; 5-23; 1995;
4. A 120 yr record of widespread contamination from mining of the Iberian pyrite belt ; Van Geen A.; Adkins J. F.; Boyle E. A.; Nelson C. H.; Palanques A.; Geology; v. 25; no. 4; p. 291–294; 1997;
5. Directiva 2006/21/CE do Parlamento Europeu e do Conselho de 15 de Março de 2006 relativa à gestão dos resíduos de indústrias extractivas e que altera a Directiva 2004/35/CE;
6. Abandoned mine sites as a source of contamination by heavy metals: A case study in a semi-arid zone; Navarro M.C., Pérez-Sirvent C., Martínez-Sánchez M.J., Vidal J., Tovar P.J., Bech J.; Journal of Geochemical Exploration 96; 183–193; 2008;
7. Morfometria e caracterização do meio físico de ambientes lacustres no Vão Do Paranã-Goiás, Brasil. Uma primeira aproximação; Carvalho T. M., Zucchi M. R.; Terra Nueva Etapa; Vol. XXV, No. 38; pp. 111-140; 2009;
8. The mineralized veins and the impact of old mine workings on the environment at Segura, central Portugal; Antunes I.M.H.R., Neiva A.M.R., Silva M.M.V.G.; Chemical Geology; 190; 417– 431; 2002;
9. Geochemistry: Exploration, Environment, Analysis; Reimann C. and Smith D. B.; Geological Society of London; v. 8; no. 3-4; p. 203-204; 2008;
10. Methodologies for Soil and Sediment Fractionation Studies; Sahuquillo A., López-Sanches J. F., Rauret, G., Ure A. M., Muntau H., Quevauviller, P. E.; Quevauviller, Ph., ed.; RSC: Cambridge, cap. 2; 2002;
11. Contributions of discharges from a historic antimony mine to metalloid content of river waters, Marlborough, New Zealand, Wilson N.J., Craw D., Hunter K.; Journal of Geochemical Exploration; 84; 127–139; 2004;
12. Trace Elements in Soils and Plants, Third Edition , Alina Kabata-Pendias, CRC PRESS, 2000;
13. Concentração de metais pesados nos sedimentos de corrente no parque estadual do itacolomi e arredores M.G.; Oliveira M.R., Roeser H. M. P. e Horn A. H.; GEONOMOS; 13(1, 2): 83-90; 2005;
14. A análise sedimentar e o conhecimento dos sistemas marinhos (versão preliminar); J. Alveirinho Dias (2004);
15. A rapid procedure for environmental sampling and evaluation of polluted sediments, Kralik M.; Universidade de Viena: Viena; 1998;
16. Contaminated sediments; Forstner U., Lecture Notes in Earth Sciences, No. 21; Springer-Verlag; 157; 1989;
17. Referência geoquímica regional para a interpretação das concentrações de Elementos químicos nos sedimentos da bacia do lago paranoá – DF; Ricardo Cosme Arraes Moreira, Geraldo Resende Boaventura, Quim. Nova, Vol. 26, No. 6, 812-820, 2003;

18. Introdution to Geochemistry; Third Edition; Krauskpopf K.B., Bird D. K.; McGraw-Hill International Editions, p. 539-540; 1995
19. Propriedades dos minerais e rochas, Geologia de Engenharia, cap. 2; 2008;
20. The effect of mining and related activities on the sediment trace element geochemistry of the Spokane River Basin, Washington, USA ; Grosbois C. A., Horowitz A. J., Smith J. J., & Elrick K. A. ; Geochemistry: Exploration, Environment, Analysis; v. 2; issue.2; p. 131-142; 2002;
21. Polluition of water and strean sediments associated with the Vale de Abrutiga uranium mine; Pinto, M. M. S. C., Silva, M. M. V. G. & Neiva, A. M. R.; Central Portugal Mine Water and the Environment; 2004, 23:66-75;
22. Arsenic sequestration by sorption processes in high-iron sediments; Robert A. Root a, Suvasis Dixit b, Kate M. Campbell c, Adam D. Jew d, Janet G. Hering c, Peggy A. O'Day; Geochimica et Cosmochimica Acta 71; 5782–5803; 2007;
23. Baselines of certain bioavailable and total heavy metal concentrations in Finland; Tarvainen T.; Kallio E.; Applied Geochemistry ; 17, 975-980; 2002,
24. Indice de Geoacumulação de Mercúrio em Sedimentos de Superfície do Estuário de Santos – Cubatão (sp); Silva W. L., Matos R. R. e Kristosch G. C.; Quim. Nova; Vol. 25; No. 5; 753-756; 2002;
25. Modeling Multiphase Reactive Transport in a Waste Rock Pile with Convective Oxygen Supply; Silva J. C., Vargas E. A., Jr., and Sracek O.; Vadose Zone J.; 8; 1038 – 1050;2009;
26. Formation of heavy metal bearing phases at a spring affected by the weathering of ore processing residues; Schubert M., Wenndich R., Weib H., Schreck P., Zeller T., Otto H. H. and Wolfram H.; European Journal of Mineralogy; 17; 119 – 128; 2005;
27. Metal Pollution in the Aquatic Environment, 2th ed.; Forstner U.; Wittman G. T. W., Springer-Verlag, with contributions by Prosi F. and Van Lierde J. H.; Springer-Verlag; Berlin; p. 399-473; 1981;
28. Prospecção Geoquímica – Princípios, Técnicas e Métodos,CPRM: Rio de Janeiro, 1998 e Rodrigues; Em PNMA II DI Subcompontente Monitoramento da Qualidade da Água; Ministério do Meio Ambiente, Licht, O. A. B., M. L. K. Brasília, 2001;
29. Environmental impact of metals derived from mining activities: Processes, predictions, prevention ; Salomons, W. Journal of Geochemicd Exploration; 52; 5-23;1995;
30. Colecta de amostras de solos, sedimentos e águas de ambientes impactados por mercúrio para monitoramento ambiental. In: Câmara, V.M.(ed.); Mercúrio em áreas de garimpos de ouro; Silva A.P. ECO/OPS; México; p. 99 – 105; 1993;
31. Sediments As Monitors of Heavy Metal Contamination In The Ave River Basin (Portugal): Multivariate Analysis of Data; Soares H.M.V.M., Boaventura R.A., Machado A.A.S.C., Esteves Da Silva J.C.G, Environmental Pollution; 105(3); 311-323;1999;
32. Characterization Of Heavy Metal Concentrations In The Sediments Of Three Freshwater Rivers In Huludao City, Northeast China; Zheng, N.; Wang, Q.; Liang, Z.; Zheng, D.; Environmental Pollution, 154: 135-142; 2008;

33. Arsenic sequestration by sorption process in highiron sediments ; Root RA, Dixit S, Campbell KM, Jew AD, Hering JG, O'Day PA; Geochimica et Cosmochimica Acta, 71:5782-5803; 2007;
34. Geochemistry and bioavailability of metals in sediments from northern San Francisco Bay; Lu X. Q.; Werner I.; Young T. M.; Environment International; Vol 31, 4, Pag. 593-602; 2005;
35. Basement-Hosted Quartz- Barite Sulfide Veins in the French Alps: A record of Alpine Tectonic Fluid Expulsion in the External Crytalline Massifs- Structural, Fluid Inclusion, and Isotope (S and Sr) Evidences; Polliand, M. & Moritz, R. Econ. Geol. Vol 94; pages: 37-56; 1999;
36. The Geochemistry of Natural Waters: Surface and Groundwater environments, 3rd ed.; Drever J. I.; Prentice Hall: New Jersey, 1997;
37. Fundamentos da ecologia da maior região de florestas tropicais; Sioli H.; Amazónia – Trad. J. Becker; Rio de Janeiro, Vozes, p. 72; 1985;
38. Fundamentos de Limnologia. 2 ed; Esteves F. A.; Interciência; p.575; 1998;
39. Análise química para avaliação da fertilidade de solos tropicais; RAIJ B. van; Andrade J.C.; Cantarella H. & Quaggio J.A.; eds.. Campinas; Instituto Agronômico; 285p; 2001;
40. Optical emission spectrometry. In: PAGE, A.L. Methods of soil analysis. Part 2. Chemical and microbiological properties. 2.ed; Soltanpour P.N.; Jones Jr J.B. & Workman S.M.; American Society of Agronomy; p.29-63; 1982,
41. Concepts, instrumentation and techniques in inductively coupled plasma optical emission spectrometry; Boss C.B. & Fredeen K.J.; New York; Perkin Elmer;110p; 1997;
42. Espectroscopias Vibracional e Electrónica; Gonsalves A.M.D'A.R., Serra M.E.S; Piñeiro M; Coimbra, Imprensa da Universidade; 2005;
43. Applications of Atomic Spectrometry to Regulatory Compliance Monitoring, 2ª ed.; Lenniss S. W., Katz S.A., Lynch R. W.; Wiley-VCH; New York; 1997;
44. Plasma spectrometry in the earth sciences: techniques, applications and future trends; Jarvis I. and Jarvis K. E.; Chemical Geology; Volume 95; Pages 1-33; 1992;
45. ICP Emission Spectrometry, A pratical Guide; Nölte J.; Willey-VCH: Weinheim, 267p;2003;
46. Effect of acid concentrations on the excitation temperature for vanadium ionic lines in inductively coupled plasma–optical emission spectrometry; Yuetsu Danzaki and Kazuaki Wagatsuma; Analytica Chimica Acta;447; 171-177; 2001;
47. Effect of the spray chamber design on steady and transient acid interferences in inductively coupled plasma atomic emission spectrometry; José-Luis Todolí and Jean-Michel Mermet J. Anal. At. Spectrom., 2000, 15, 863-867;
48. Análise Química Quantitativa, 6ª edição; Vogel A.R.; Mendham J.; Denney R.C.; Barnes, J.D.; Thomas M.; Editora LTC, 462p; 2002;
49. Instrumentation and Tecniques in Industrively Coupled Plasma Optical Emission Spectrometry; Boss C., Fredeen K., Concept; PerkinElmer: USA; 1997;

50. Aerosol desolvation studies with a thermospray nebulizer coupled to inductively coupled plasma atomic emission spectrometry; Mora J., Todoli J., Rico I., Canals A.; Analyst; 123; 1226; 1998;
51. Factorial design for multivariate optimization of an on-line preconcentration system for platinum determination by ultrasonic nebulization coupled to inductively coupled plasma optical emission spectrometry; Cerutti S., Salonia J. A., Ferreira S. L, Olsina R., Martinez L.D. ; Talanta; 63; 1077; 2004;
52. Avaliação das Condições Operacionais de Espectroscopia de Emissão Óptica com Plasma Acoplado Indutivamente com Configuração Axial; Trevizan L. C.; Tese de Doutoramento Química Analítica ; 2007;
53. Colorimetria. In Infopédia [Em linha]. Porto: Porto Editora, 2003-2010. [Consult. 2010-08-20];
54. Analytical methods for pesticides, plant growth and food additives; Knüsli, E. A.; Academic Press, in Zweig; p.33-36; 1964;
55. Colorimetric methods for the determination of simazine and related chloro-s-triazines; Ragab M. T. H., McCollum J. P.; Journal of Agricultural Food Chemestry; v. 16; n.2; 1968;
56. Chatanika/TRIAD observations of unstable ionization enhancements in the auroral F-region Vickrey J.F., Rino C.L. and Potemra T.A.; Geophysical Research Letters 7, pp. 789–792; 1980;
57. Análisis de minerales y elementos traza en alimentos; Kastenmayer, P.; SS Nielsen - Introduction to the Chemical Análisis of foods; p.271-294; 1994;
58. Greenfield H., Southgate D. A. T.; Food composition data: Production, management and use. 2ed. Food and Agriculture Organization of United Nations (FAO), Rome, 2003;
59. Validation d'une methode de dosage. Aplication a la'analyse dês amines biogénes du vin; Monteiro M.J.P., Bertrand A.; Feuillet Vert de l'OIV; 970; 1994;
60. Quality assurance of chemical measurements; Taylor J.K.; CRC Press, New York; 1987;
61. Norma ISO 8466-1 Water Quality – Calibration and Evaluation of Analytical Methods and Estimation of performance characteristics, Part 1: Statistical Evaluation of the linear calibration function (http://www.iso.org/iso/catalogue_detail.htm?csnumber=15664);
62. Norma ISO 8466-2 Water Quality – Calibration and Evaluation of Analytical Methods and Estimation of performance characteristics, Part 2: Calibration Strategy for non-linear second-order calibration function (http://www.iso.org/iso/catalogue_detail.htm? csnumber=34816);
63. On Lines and Planes of Closest Fit to Systems of Points in Space"; Pearson K.; Philosophical Magazine 2 (6); 559–572; 1901;
64. Uber linear Methoden in der Wahrschein-lichkeitsrechung, Annales Academiae; Karhunen K.; Scientiarum Fennicae; Series A, 137; 1947;
65. Extrended Q model- Objective definition of external end members in the analysis of mixtures; Full W., Ehrlich R. and Klovan J.; Mathematical Geology; 13; 331-334; 1981;
66. Recursive vs. nonrecursive systems: An attempt at synthesis; Strotz R. H. and Wold H. O. A.;, Econometrica; 28, 417-427; 1960;
67. Self modeling curve resolution; W. Lawton and E. Sylestre, Technometrics; 13; 617-633; 1971;

68. Chemometrics and Intelligent Laboratory Systems; Elsevier Science Publishers B. V., Amestardam; 2; 37-52; 1987;
69. New Rontes from Minimal Approximation Errol to Principal Components; Miranda A. A., Le Borgne Y. A. and Bontempi G.; Neural Processing Letters, Spvringer; volume 27; Number 3; 2008;
70. A Tutorial on principal component analyses; Shlens J.; Derivation, Discussion and Singular Value Decomposition; Version 1; 25 March 2003;
71. Multivariate data analysis. 5.ed. HAIR J.F. et al.; Englewood Cliffs: Prentice-Hall; 1998. 730p
72. Avaliação do uso de métodos quimiométricos em análise de solos; Sena M. M., Poppi R.J., Frighetto R.T.S. e Valarini P.J.; Quimica Nova, 23(4); 547-556; 2000;
73. An introduction to partial least squares regression; Tobias D. Randall; SAS Institute Inc., Cary, NC; 2008;
74. "Statistics for analytical chemistry"; J.C. Miller, J.N. Miller; Ellis Horwood, New York; 1988;
75. NIST/Sematech, "Engineering Statistics Handbook", ttp://www.itl.nist.gov/div898/handbook;
76. Concentração de metais pesados nos sedimentos de Corrente no parque estadual do Itacolomi e Arredores, mg; Oliveira, M. R., Roeser H.M.P., Horn A. H.; GEONOMOS; 13(1, 2); 83-90; 2005;
77. Estudo Da Mobilização De Metais E Elementos Traços Em Ambientes Aquáticos Do Semiárido Brasileiro Aplicando Análises De Componentes Principais; Santos J. S. e Santos M. L. P; Quim. Nova; Vol. 31; No. 5; 1107-1111; 2008;
78. Prospecção na área de Sarzedas-Castelo Branco, zona de Pomar-Galdins. Informação de Proposta de Sondagens; Viegas L., Santarem R., Rodrigues L., Moreira J.; Serviço de Fomento Mineiro, Portugal, 19 pp.; 1987;
79. Análise química Quantitative; Harris D. C.; San Francisco: W.H. Freeman; 2003;
80. Estratégias para Aplicação no Trabalho do Aprendido em Treinamento: Proposição Conceitual e Desenvolvimento de uma Medida; Pilati R., Andrade J. E. B.; Psicologia: Reflexão e Crítica, 18(2) pp.207-214; 2005;
81. VISÃO COMPUTACIONAL, MEI/1; Proença H. P.; Universidade da Beira Interior, Departamento de Informática; 2008/2009;
82. www.mathworks.com
83. www.sia.com.br/mathematica.htm)
84. www.gnu.org/software/octave/index.html
85. www.scipy.org

Anexos

Anexo A1

Tabela A1.1 – Valores críticos para teste de Grubbs (_ = 0:05) [ISO, ASTM E-178]

n	3	4	5	6	7	8	9
G	1.153	1.463	1.672	1.822	1.938	2.032	2.110
n	10	15	20	25	50	100	
G	2.176	2.409	2.557	2.663	2.956	3.207	

NOTA: Os valores de Grubbs (α= 0.05) também podem ser estimados com base na expressão:

$$G_\alpha = \left(\frac{N-1}{\sqrt{N}}\right)\sqrt{\frac{t^2_{\alpha(N-2)}}{(N-2)+t^2_{\alpha(N-2)}}}$$

Tabela A1.2: Valores críticos da distribuição t-student bilateral. Os valores referentes à distribuição unilateral devem ser consultados através da coluna referente a 2α (ex: teste a 5% de uma cauda -> α = 2x0:05 = 0:10).

	α					α					α			
ν	0.10	0.05	0.02	0.01	ν	0.10	0.05	0.02	0.01	ν	0.10	0.05	0.02	0.01
2	2.92	4.30	6.96	9.92	12	1.78	2.18	2.68	3.05	24	1.71	2.06	2.49	2.80
3	2.35	3.18	4.54	5.84	13	1.77	2.16	2.65	3.01	26	1.71	2.06	2.48	2.78
4	2.13	2.78	3.75	4.60	14	1.76	2.14	2.62	2.98	30	1.70	2.04	2.46	2.75
5	2.02	2.57	3.36	4.03	15	1.75	2.13	2.60	2.95	35	1.69	2.03	2.44	2.72
6	1.94	2.45	3.14	3.71	16	1.75	2.12	2.58	2.92	40	1.68	2.02	2.42	2.70
7	1.89	2.36	3.00	3.50	17	1.74	2.11	2.57	2.90	50	1.68	2.01	2.40	2.68
8	1.86	2.31	2.90	3.36	18	1.73	2.10	2.55	2.88	60	1.67	2.00	2.39	2.66
9	1.83	2.26	2.82	3.25	19	1.73	2.09	2.54	2.86	80	1.66	1.99	2.37	2.64
10	1.81	2.23	2.76	3.17	20	1.72	2.09	2.53	2.85	100	1.66	1.98	2.36	2.63
11	1.80	2.20	2.72	3.11	22	1.72	2.07	2.51	2.82	∞	1.64	1.96	2.33	2.58

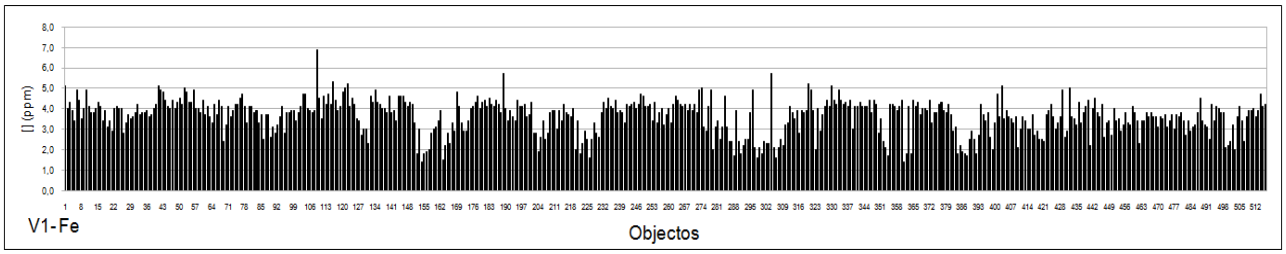
Tabela A1.3 – Valores críticos da distribuição de Fisher-Snedcor unilateral (α= 0:05)

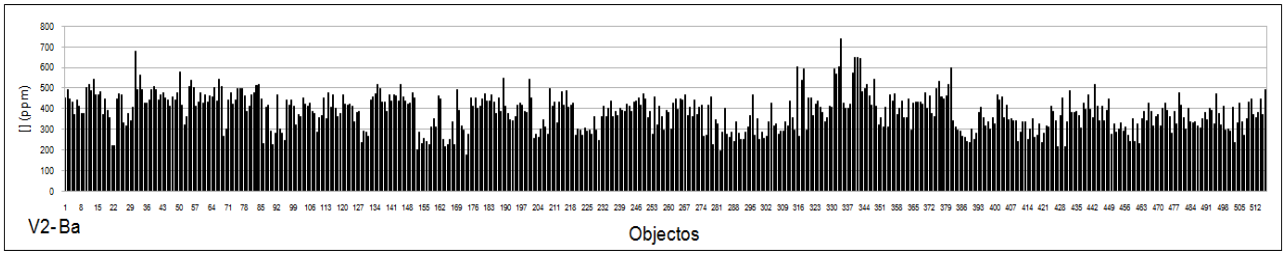
$\nu_1 \backslash \nu_2$	1	2	3	4	5	6	7	8	9	10	12	15	20
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.703	8.660
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619	4.558
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938	3.874
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.218	3.150
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006	2.936
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845	2.774
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.719	2.646
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.533	2.459
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.463	2.388
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403	2.328
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.352	2.276
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.308	2.230
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.269	2.191
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.234	2.155
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.151	2.071
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.089	2.007
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.015	1.932

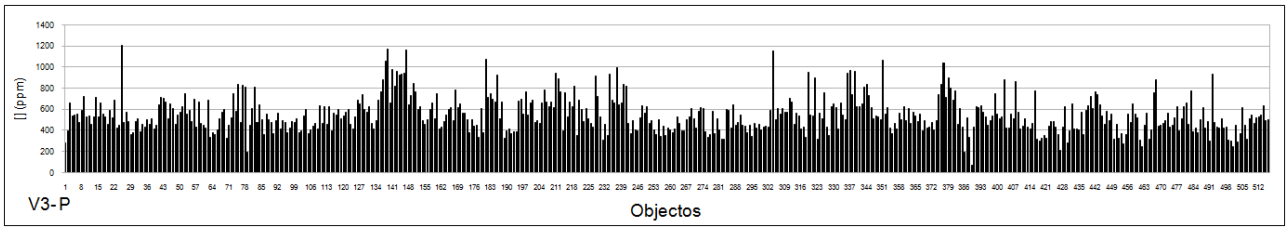
Tabela A1.4 – Valores críticos da distribuição de Fisher-Snedcor unilateral (α = 0:01)

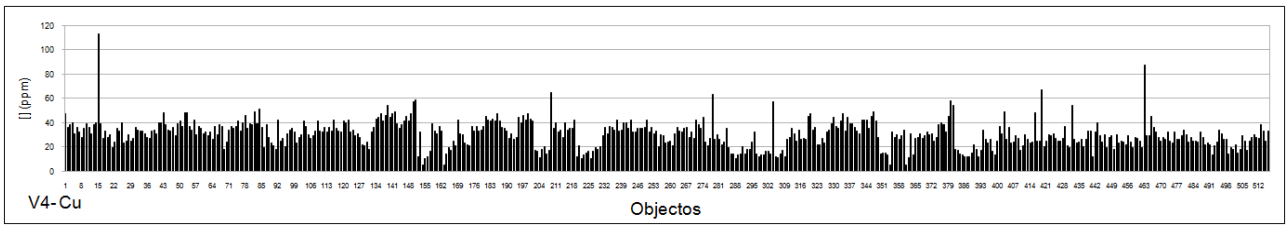
	1	2	3	4	5	6	7	8	9	10	12	15	20
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55

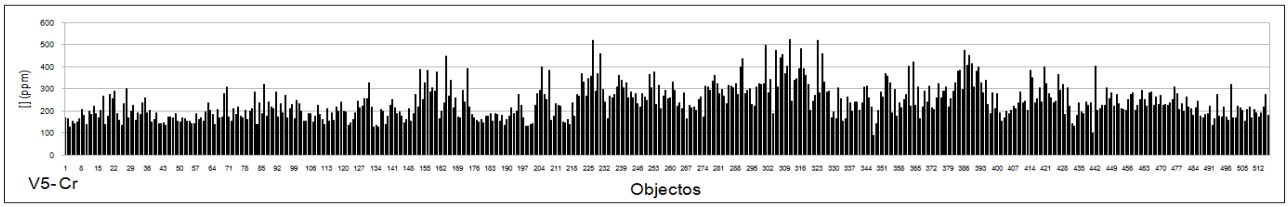
Anexo A2

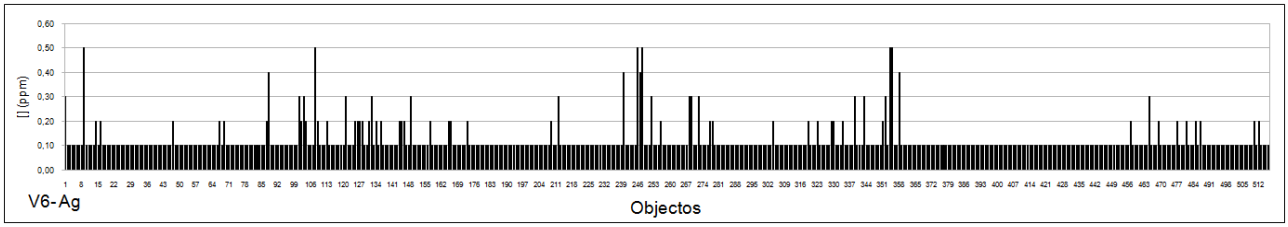

[] (ppm)
V1-Fe
Objectos

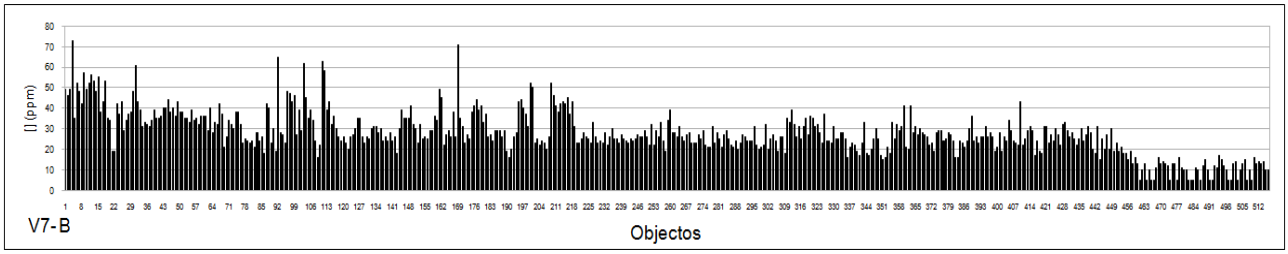

[] (ppm)
V2-Ba
Objectos

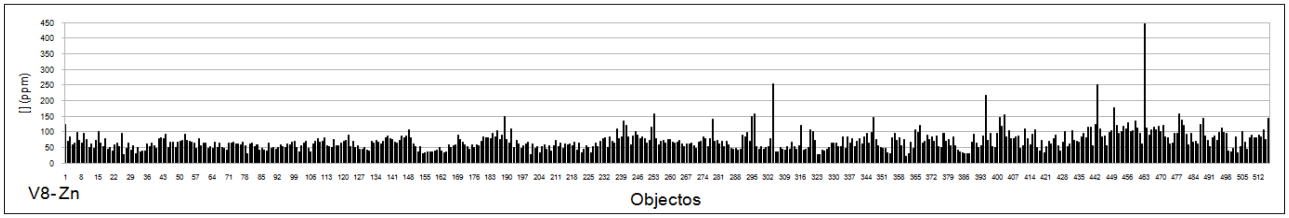

[] (ppm)
V3-P
Objectos

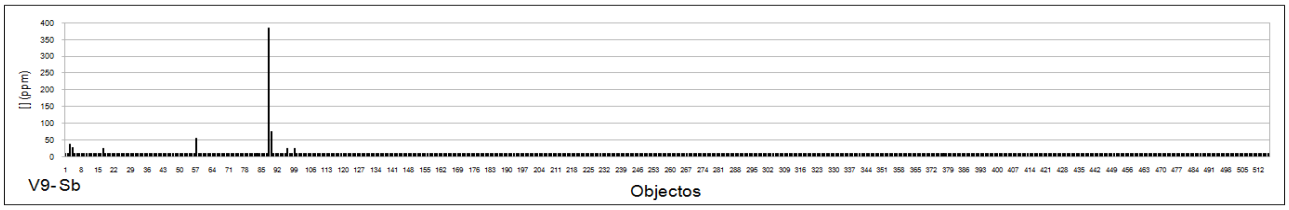

[] (ppm)
V4-Cu
Objectos

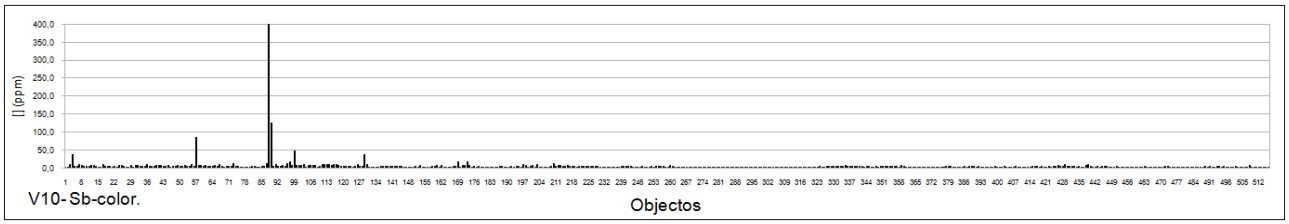

[] (ppm)
V5-Cr
Objectos

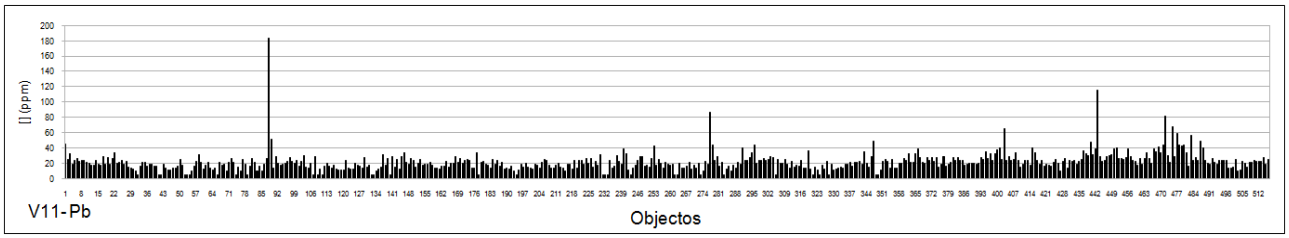

[] (ppm)
V6-Ag
Objectos

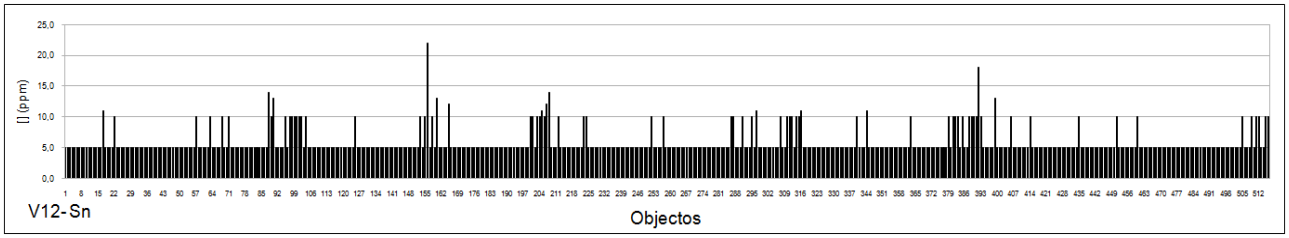

[] (ppm)
V7-B
Objectos

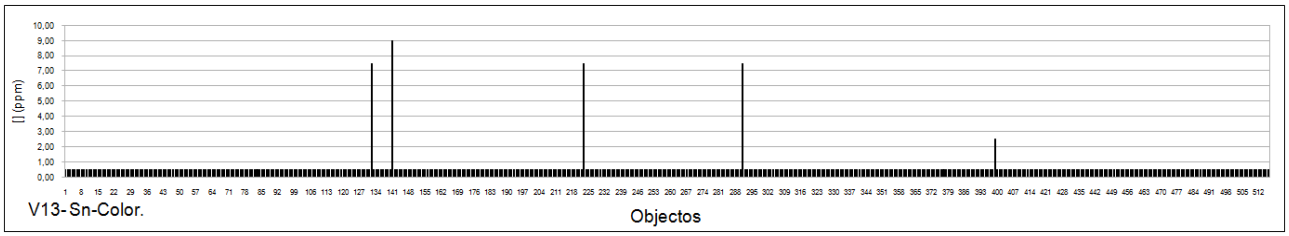

[] (ppm)
V8-Zn
Objectos

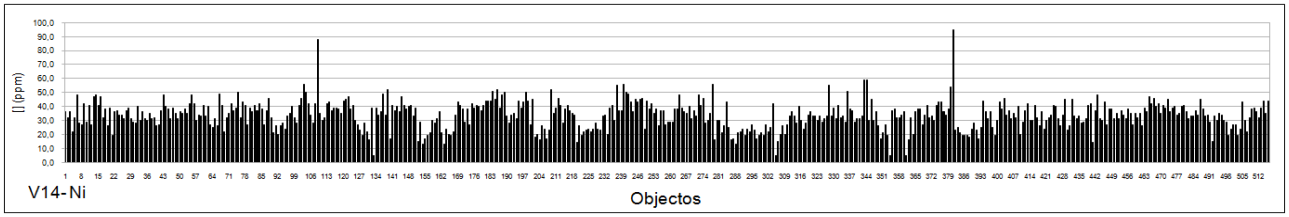

[] (ppm)
V9-Sb
Objectos

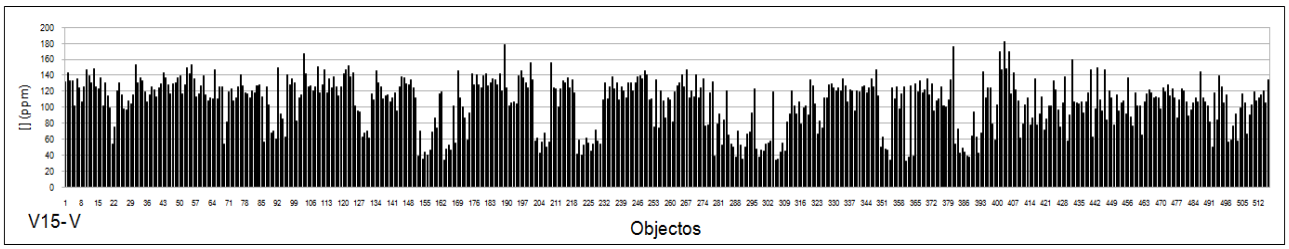

[] (ppm)
V10-Sb-color.
Objectos

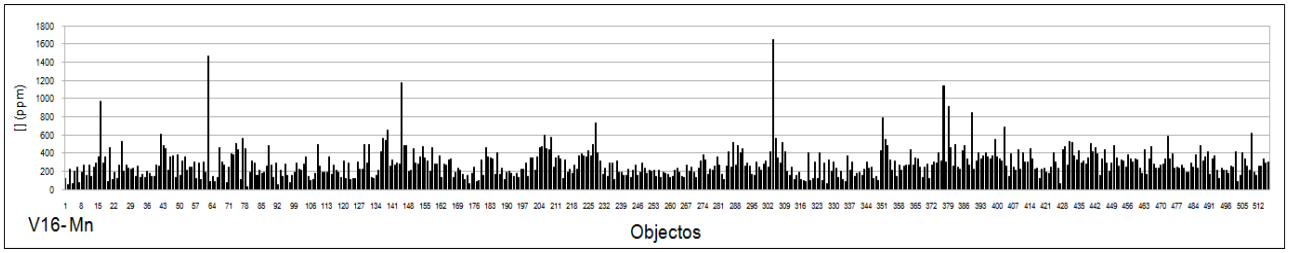

[] (ppm)
V11-Pb
Objectos

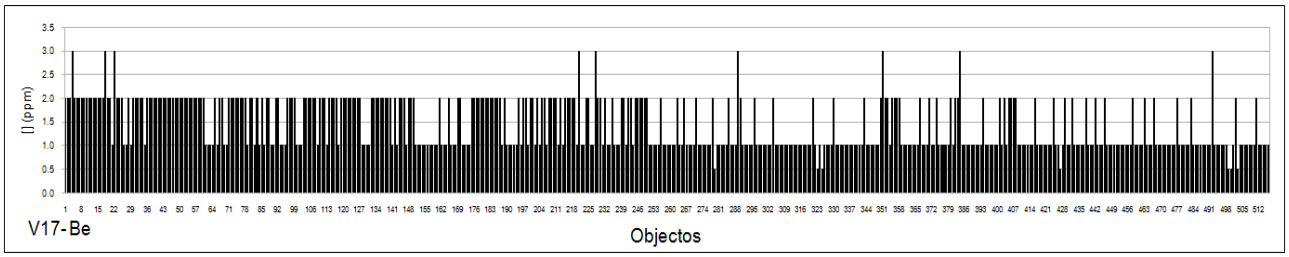

[] (ppm)
V12-Sn
Objectos

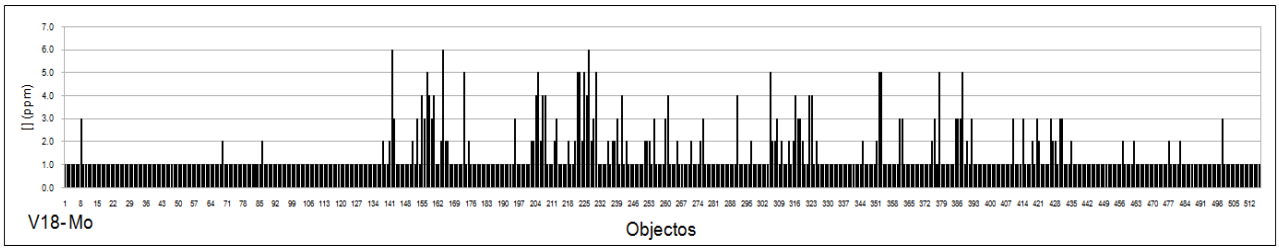

[] (ppm)
V13-Sn-Color.
Objectos

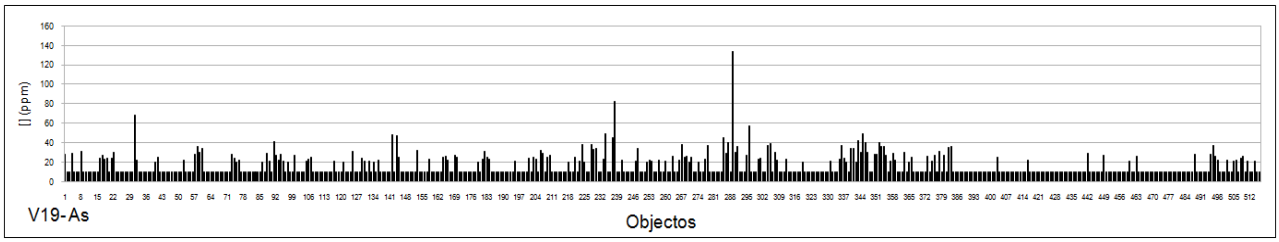

[] (ppm)
V14-Ni
Objectos

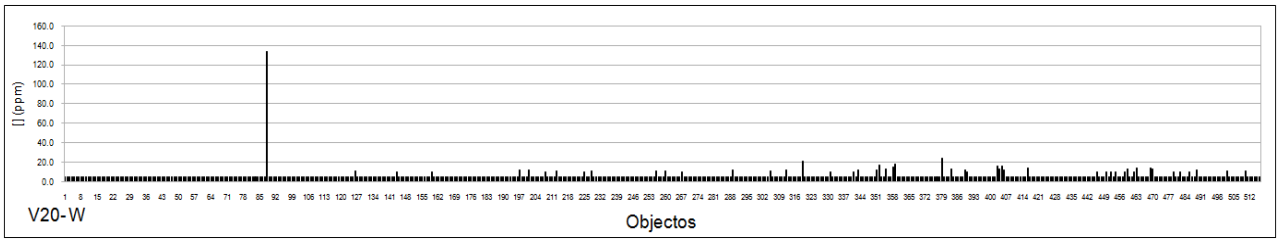

[] (ppm)
V15-V
Objectos

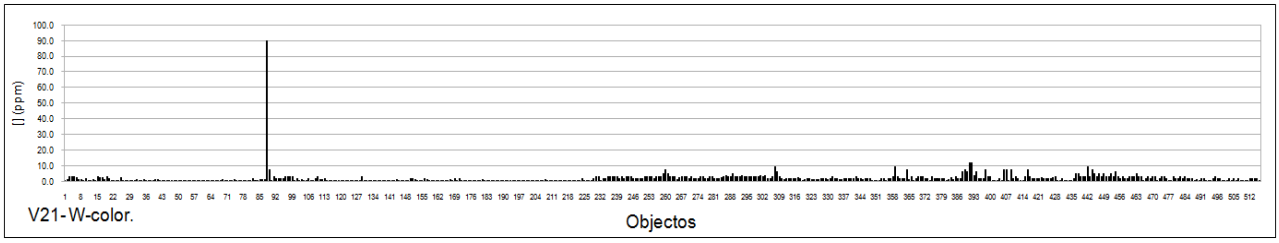

[] (ppm)
V16-Mn
Objectos

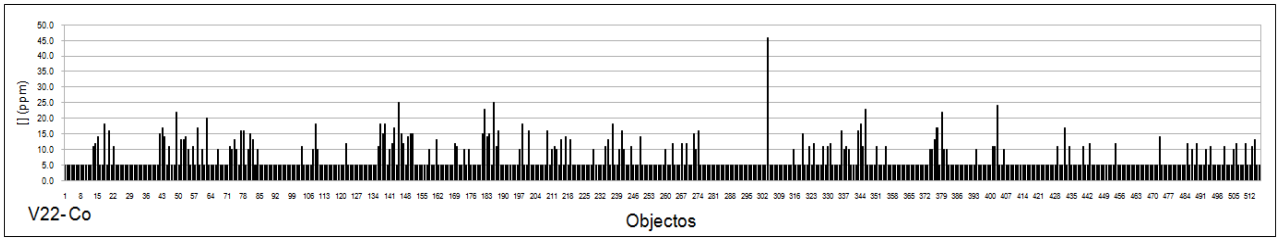

[] (ppm)
V17-Be
Objectos


[] (ppm)
V18-Mo
Objectos


[] (ppm)
V19-As
Objectos


[] (ppm)
V20-W
Objectos


[] (ppm)
V21-W-color.
Objectos


[] (ppm)
V22-Co
Objectos

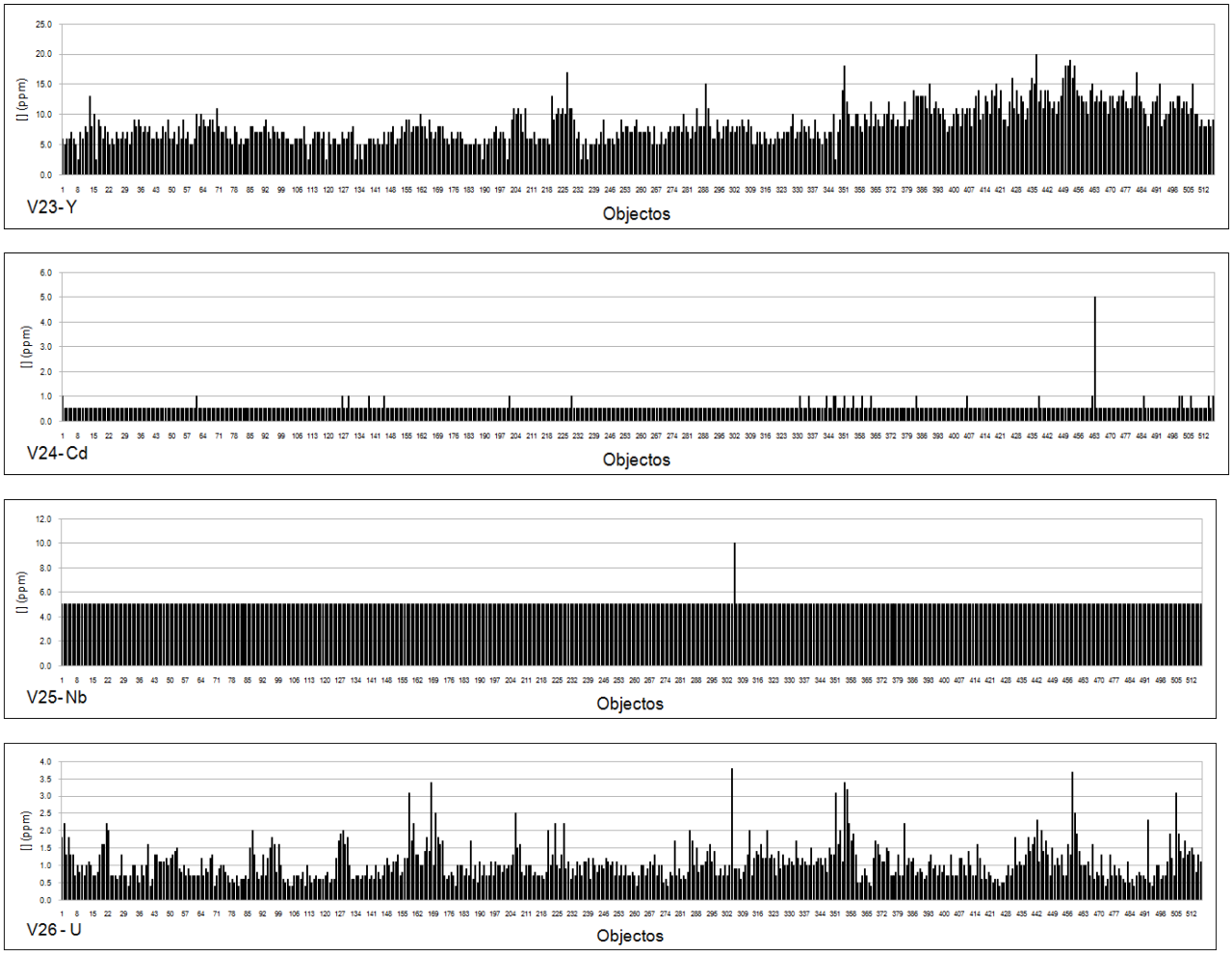
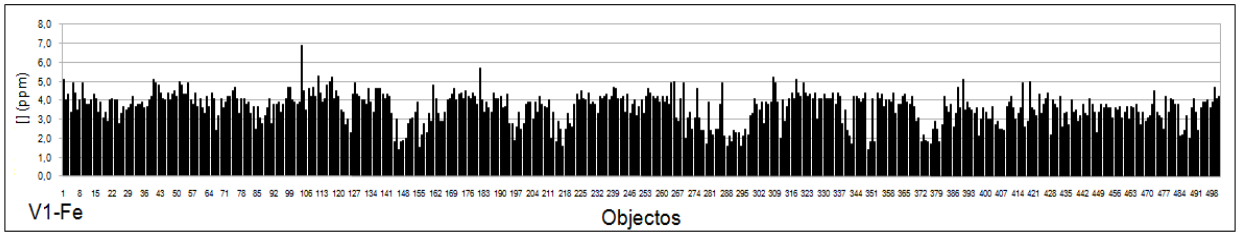
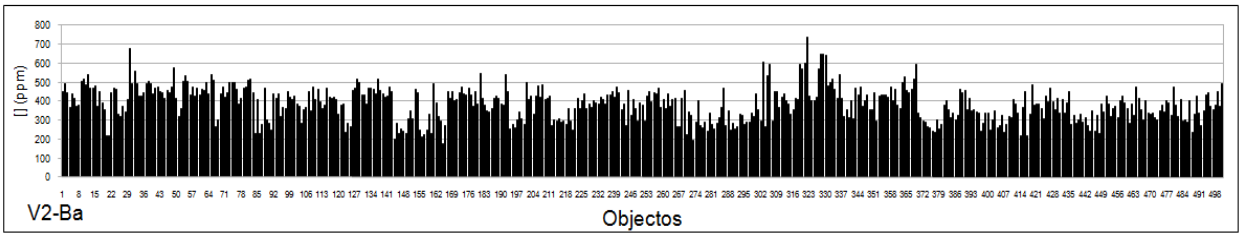
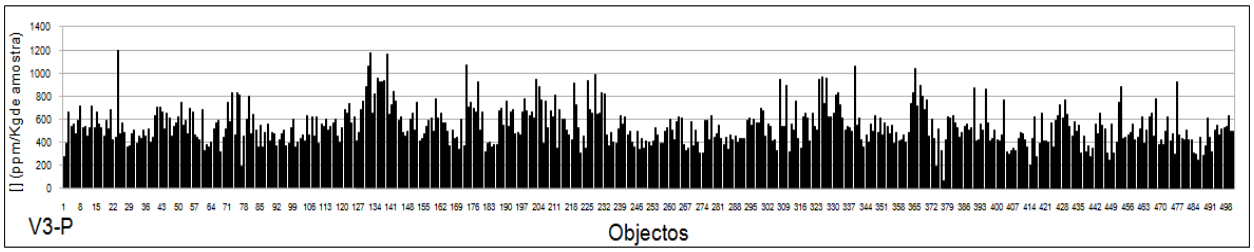
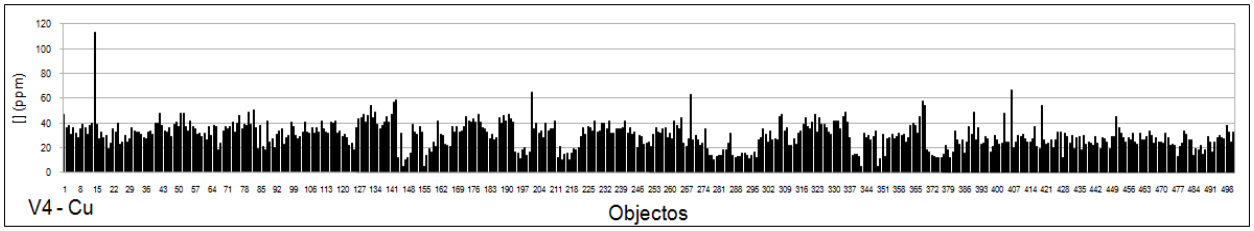


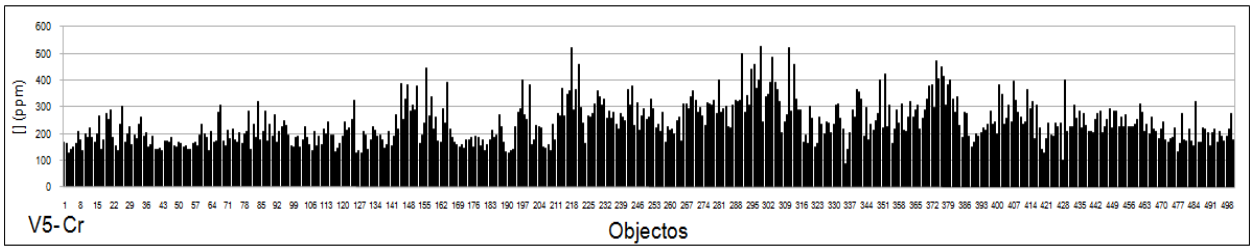
Figura A2.1 – Representação gráfica das 26 variáveis em estudo.

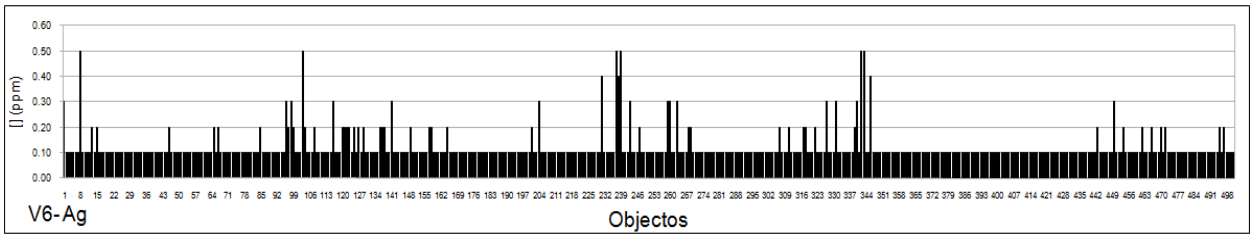

[] (ppm)
V1-Fe
Objectos

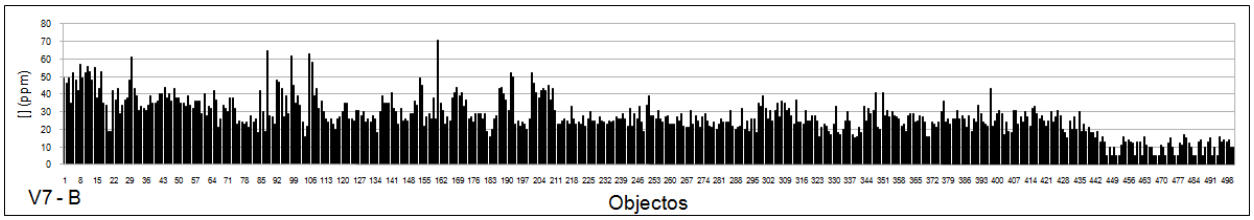

[] (ppm)
V2-Ba
Objectos

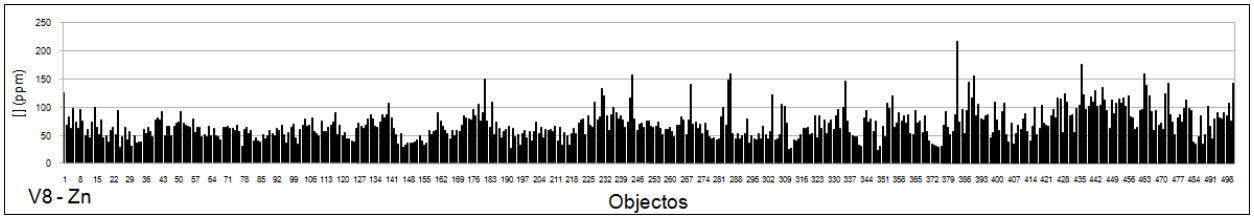

[] (ppm/Kg de amostra)
V3-P
Objectos

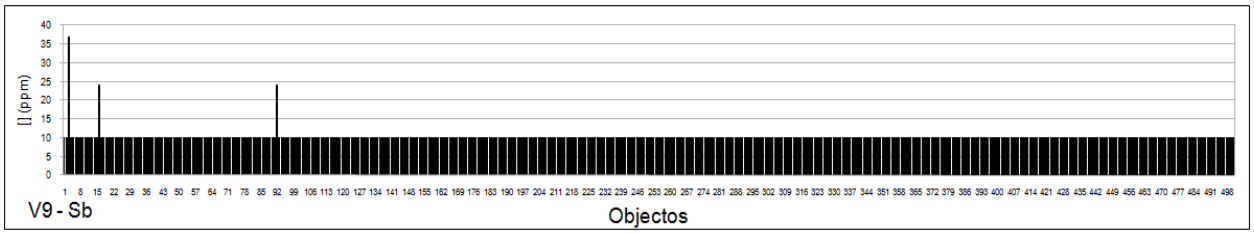

[] (ppm)
V4 - Cu
Objectos

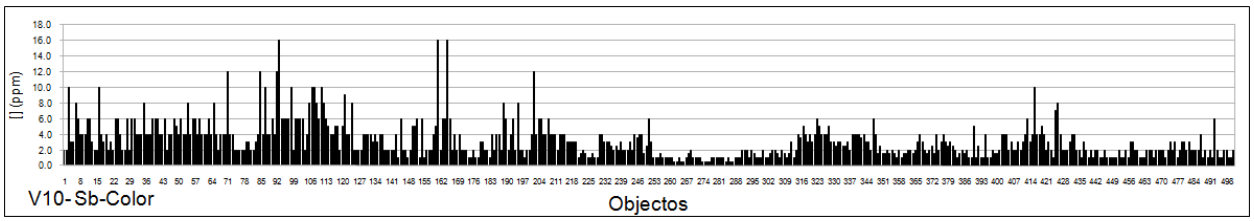

[] (ppm)
V5- Cr
Objectos

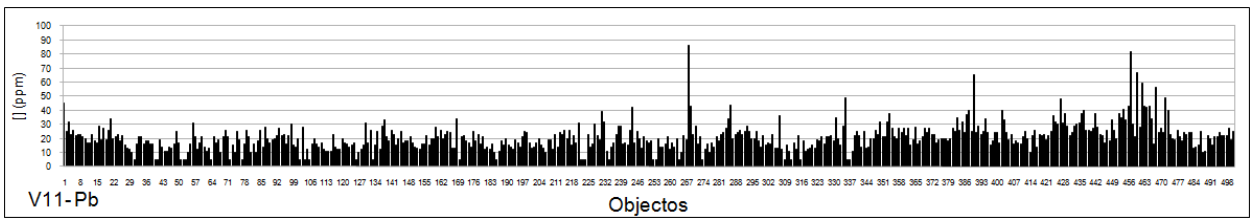

[] (ppm)
V6- Ag
Objectos

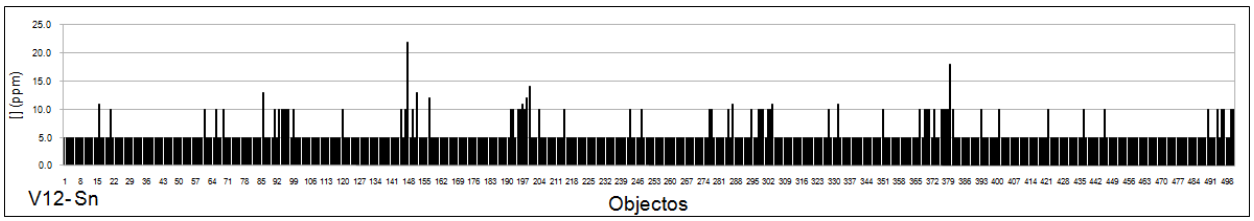

[] (ppm)
V7 - B
Objectos

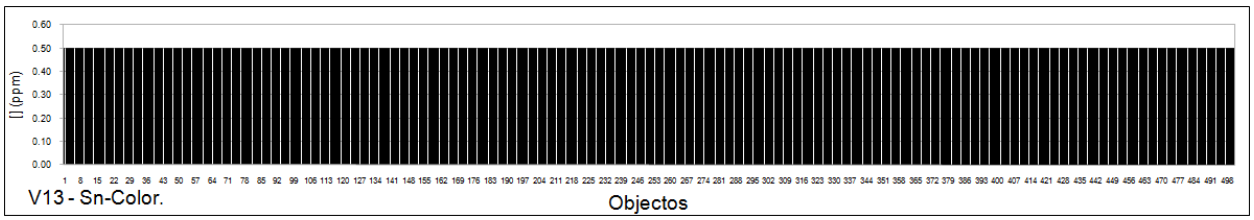

[] (ppm)
V8 - Zn
Objectos

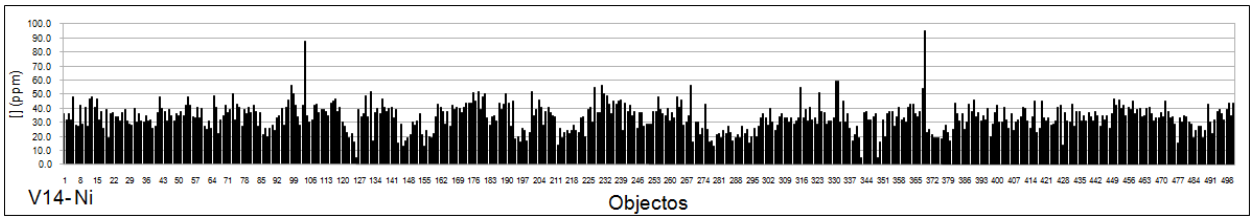

[] (ppm)
V9 - Sb
Objectos

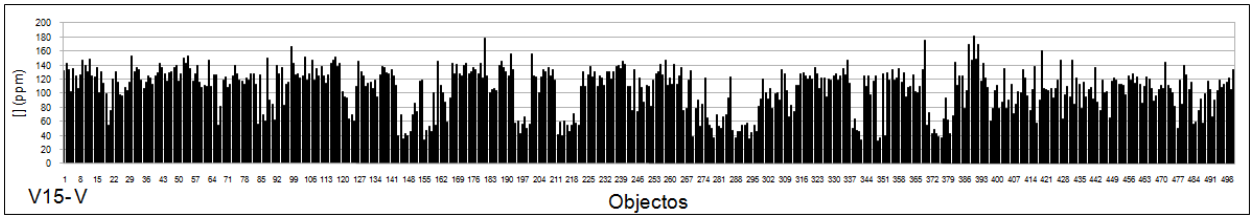

[] (ppm)
V10- Sb-Color
Objectos

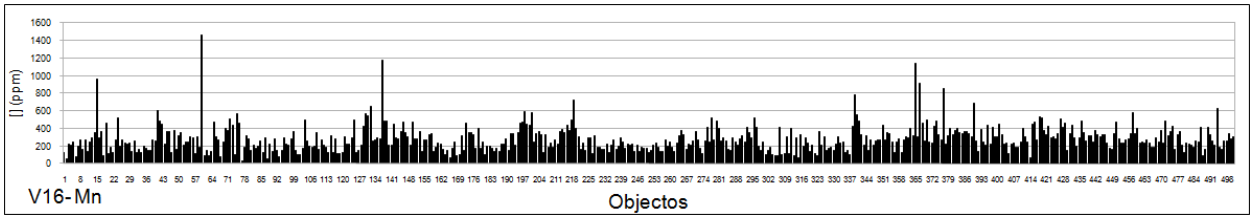

[] (ppm)
V11- Pb
Objectos

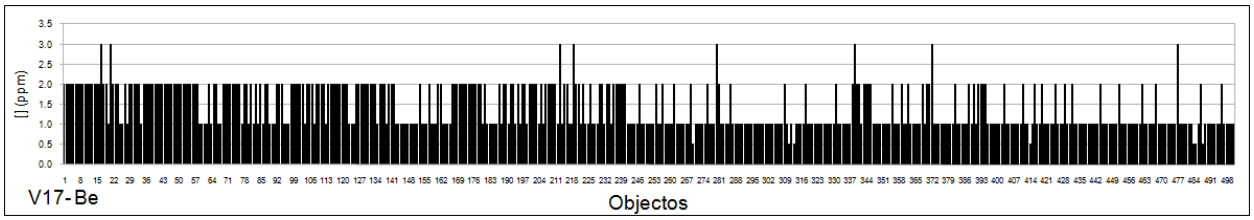

[] (ppm)
V12- Sn
Objectos

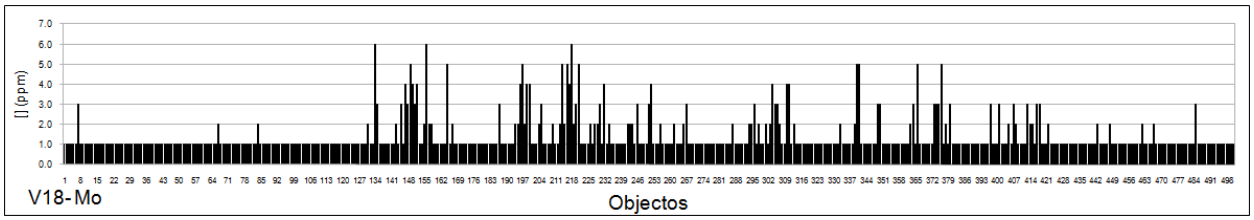

[] (ppm)
V13 - Sn-Color.
Objectos

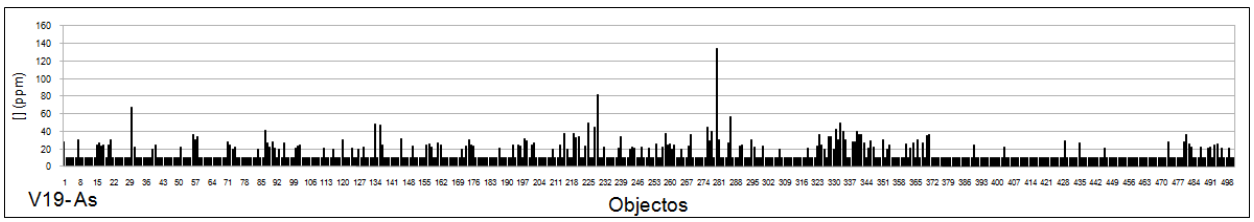

[] (ppm)
V14- Ni
Objectos

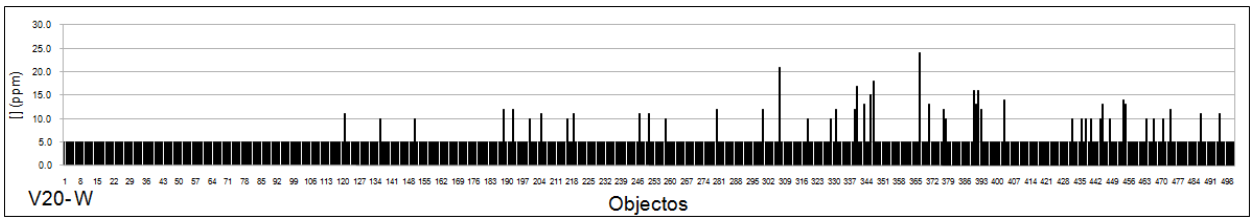

[] (ppm)
V15- V
Objectos

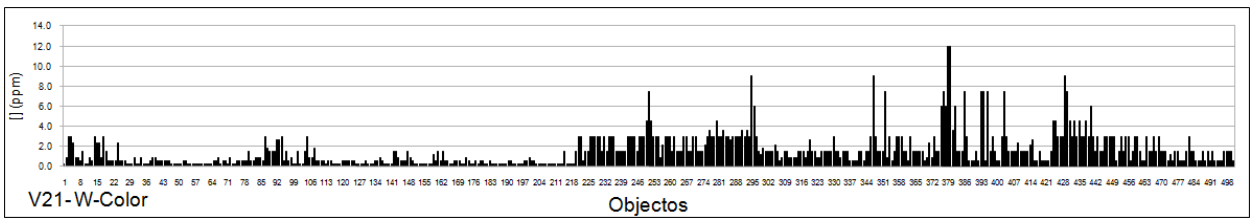

[] (ppm)
V16- Mn
Objectos

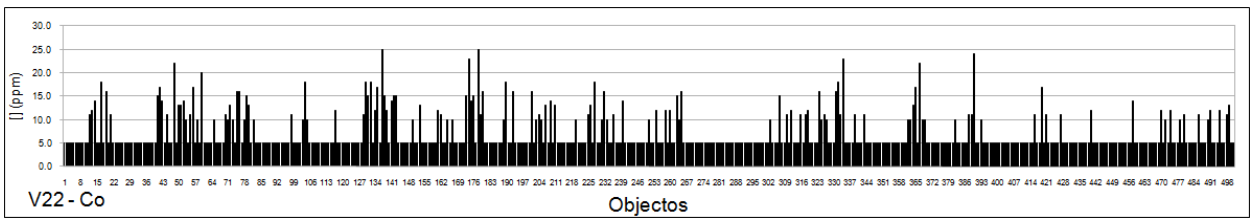

[] (ppm)
V17- Be
Objectos

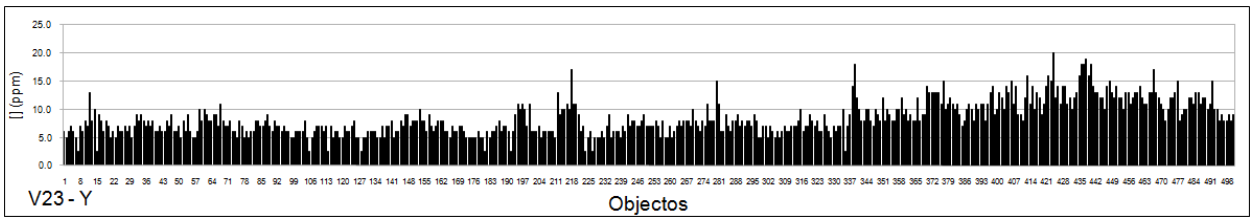

[] (ppm)
V18- Mo
Objectos


[] (ppm)
V19- As
Objectos


[] (ppm)
V20- W
Objectos


[] (ppm)
V21- W-Color
Objectos


[] (ppm)
V22 - Co
Objectos


[] (ppm)
V23 - Y
Objectos

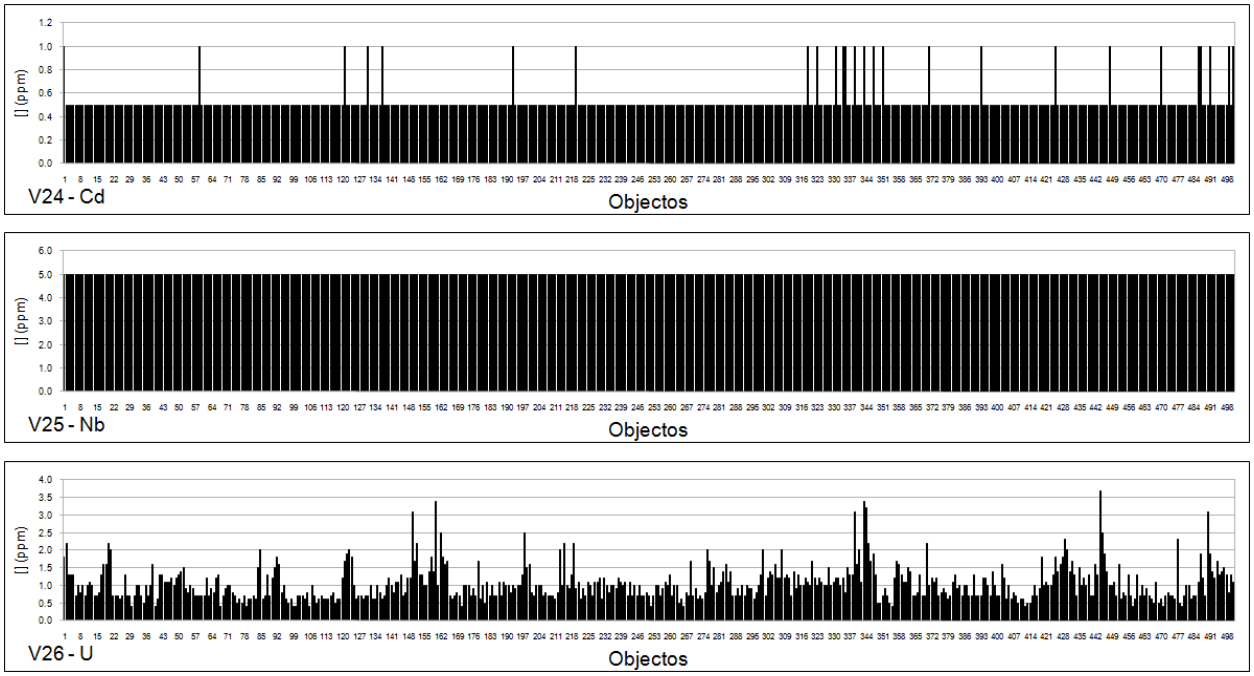


Figura A2.2 - Representação das variáveis em estudo após terem sido removidos os outliers.