UNIVERSITÀ DEGLI STUDI DI MILANO
SCUOLA DI DOTTORATO
TERRA, AMBIENTE E BIODIVERSITÀ

Facoltà di Scienze Matematiche, Fisiche e Naturali
Dipartimento di Scienze della Terra "Ardito Desio"

Dottorato di Ricerca in Scienze della Terra
Ciclo XXIV – Settore scientifico-disciplinare GEO/08

Source identification of environmental pollutants using chemical analysis and Positive Matrix Factorization

Ph.D. Thesis

Comero Sara
Matr. N. R08047

Tutori		*Coordinatore*
Prof. Luisa De Capitani	**Anno Accademico**	**Prof. Elisabetta Erba**
Dott. Bernd Manfred Gawlik	**2010-2011**	

A Marco
e a Sunny, senza dubbio

Table of contents

Abstract

Multivariate modeling techniques are successfully used in different areas of environmental research because of their ability to process large data sets. The main objective of their application lies in the determination of data structures and hidden information which account for the data set variability.

This thesis work seeks to explore the application of the positive matrix factorization (PMF) technique to different geochemical data sets on three spatial scales: local, pan-regional and pan-European. In particular, we focus on PMF identification of pollutants/contamination sources (e.g., anthropogenic and natural pollution) and chemical/physical processes (e.g., mineralization, weathering and corrosion) characterizing the data sets under examination.

PMF analysis was carried out on four data sets with different spatial scale:

- at local scale, geochemical characteristics of soil samples at the abandoned *Coren del Cucì* mine dump were examined. A GIS-based approach was also combined with PMF results for a better source resolution. Five factors were determined: (i) two geo-morphological backgrounds characteristic of the area outside the dump; (ii) a source of mineralization situated inside the waste disposal area; and (iii) two different geochemical anomaly zones;
- at a national level, eleven alpine lakes site in the Northern Italy were considered. X-ray fluorescence analyses on lake sediments were evaluated by PMF. Four interpretable mineralogical/chemical features were identified: (i) phosphate and sulphur source; (ii) carbonates; (iii) silicates; and (iv) heavy metal-bearing minerals. Also, to properly modify input information, a new PMF factor was determined, explaining a possible Pb contamination source;
- in the pan-regional context, sediments of the Danube River basin, which cover an area of 817.000 km^2, flowing through nine European countries, were analysed. The objective was to draw out information about the natural vs. anthropogenic origin of heavy metals and to determine the role of tributaries. Three factors were identified: (i) a carbonate component characterized by Ca and Mg; (ii) an alumino-silicate component dominated by Si and Al content and the presence of some metals attributed to natural processes; (iii) an anthropogenic source identified by Hg, S, P and some heavy metals load. Considering

only the tributaries input, an additional source probably attributed to the use of fertilizers in agriculture was determined;

- finally, a pan-European data set comprising sewage sludge from European waste water treatment plants was obtained. The final objective was to link the silver content to the increasingly use of silver nanoparticles in a variety of house-hold and personal care products. Here, method validation procedure was applied to the measured elements in order to compute correct uncertainties to be used in PMF application. The four resulting factors could be described by: (i) copper dissolution from water pipe lines; (ii) engineered silver nanoparticles load; (iii) anthropogenic influence suggested by the presence of different metals; and (iv) iron variation due to the use of this element for phosphorus removal in sewage sludge.

These studies provide first evidence that PMF could be successfully applied to geochemical data sets at different spatial scale.

Chapter 1

Multivariate modelling

1.1. Introduction

Multivariate statistical techniques have been widely used in different branches of environmental research (Kaplunovsky, 2005; Viana *et al*., 2008; Mostert *et al*., 2010) because they provide a useful tool for the analysis of large data sets. The concept of '*multivariate'* deals with the statistical analysis of data sets which contains more than one variable. The main objective of their application is to reduce the dimensionality of examined data sets but also to point out any trend and/or correlation among variables.

In particular, when the application of multivariate statistical methods is addressed to the identification and quantification of natural/anthropogenic sources, they are generally termed *receptor models* (Gordon, 1988). Receptor modelling is based on the information registered at the impact point, the receptor, which is usually given by concentrations of chemicals measured at the sampling location (Hopke, 2003). In this way, they are complementary to source-oriented dispersion models (prognostic models) which are based on sources emission inventory to estimate concentrations measured at receptors.

1.2. Types of receptor models

Depending on the type of information at the receptors, receptor models divide in two main branches: *chemical mass balance models* (CMB) and *multivariate receptor models* (Pollice A., 2009). In the first case, main sources number and their composition profiles must be known *a priori*, while multivariate receptor models assume only the knowledge of observations (usually chemical concentrations) at the receptor sites. However, as reported in **Fig. 1,** they represent two extremes.

Most commonly used receptor model in physical and chemical sciences applications are Cluster Analysis (CA), Principal Component Analysis (PCA), Unmix, Target Transformation Factor Analysis (TTFA) and Positive Matrix Factorization (PMF). However, in geochemical studies many investigators prefer the use of PCA and CA, probably due for their ease to use and

availability in major statistical software packages. Only in recent years the applicability of other techniques has been tested in soils, sediments and water compartments (Bzdusek et al., 2006; Lu et al., 2008; Huang and Conte, 2009).

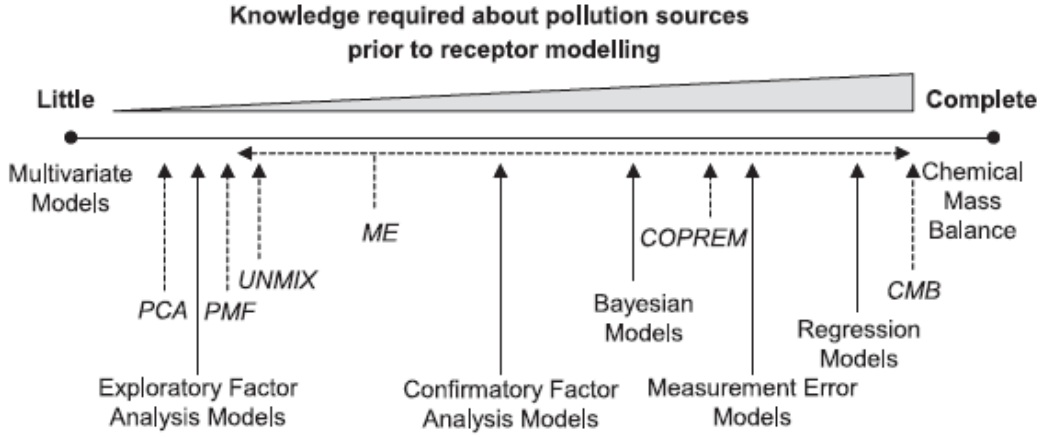


Fig. 1: types of receptor models ordered basing on the knowledge about the source prior to the modeling (from Viana *et al*., 2008)

In this thesis we want to draw the attention on the PMF approach. The reason lies in its property to be a non data-sensitive technique; no pre-treatment (e.g. data normalization and/or standardization) of data is necessary. Moreover, incorporating the variable uncertainties in the resolving algorithm, problematic data such as below-detection-limit and outliers can be appropriately weighted.

The final objective was to explore the PMF applicability in environmental data sets characterized by a different the spatial scale. In the last application, the method validation technique was also performed in order to determine laboratory data uncertainties to be use in the statistical technique.

Chapter 2

Cluster analysis

2.1. Introduction

Cluster analysis is a multivariate pattern recognition technique that helps to identify natural groups of classes existing in data sets (Hardle and Simar, 2003).

It has been widely used in environmental studies (Swanson *et al*., 2001; Treffeisen *et al*., 2004; Helstrup *et al*., 2007; Dragović and Mihailović, 2009); in particular, it garnered widespread interest in geochemical applications (Grande *et al*., 2003; Templ *et al*.2008; Ribeiro *et al*., 2010; Morrison *et al*., 2011). In example, it was applied to classify variables on the basis of the similarities of their geochemical properties (Yongming *et al*., 2006; Bhuiyan *et al*., 2010), but also to identify the chemical relationships between samples showing similar chemical characteristics (Helstrup *et al*., 2007).

Cluster analysis can be performed using different clustering algorithms (some of them are listed in § 2.3). Prior to classification criteria, a distance measure (cfr. § 2.2) must be defined, which determines the similarity or dissimilarity between samples or variables.

Cluster analysis is a data-sensitive technique and usually requires a previous univariate analysis of the data set (Reimann *et al*., 2002). In fact, geochemical datasets are often characterized by heavily skewed distributions and normalization procedures have to be applied to obtain a more symmetric distribution (Webster, 2001). Usually log-transformation and Box-Cox, explained in § 2.4, are used.

Moreover, geochemical data set are characterized by variables which show a high variation in concentrations values. In example, data sets often consist of concentrations of major, minor and trace elements, which can vary over orders of magnitude. This can produce inappropriate cluster analysis results because, if the clustering method is based on distance coefficients, outputs are more strongly influenced by the variable which shows the greatest magnitude (Templ *et al*., 2008). In these cases, additional standardization techniques, explained in § 2.5, are necessary prior to the cluster analysis.

Problematic data such as outliers, if not identified, can give incorrect cluster results. Although they may contain important information, i.e. they may be indicative of mineralization (Filzmoser

et al, 2005) they should be removed from the data set. One method to identify them, which was used in this thesis, is based on the Mahalanobis distance (Filzmoser *et al*, 2005).

Variables with a high proportion of observations below the detection limit are usually omitted from cluster analysis. In fact, substituting them with appropriate estimated, usually ½ the detection limit, can significantly alter the clustering determination (Templ *et al*., 2008). During the PhD work, variables with more than 5% of BDL have been omitted from cluster analysis.

Cluster analysis was here applied to two geochemical data sets characterized by a strong skewness. In both applications the Ward agglomerative hierarchic method and the Euclidean distance were used. These classification criteria are the more frequent choices in geochemical applications (Zupan *et al*., 2000; Helstrup *et al*., 2007). In the *Coren del Cucì* mine site application (Ch. 6) CA was performed to group sampling locations in order to extract more homogeneous sub-populations for further data analysis, i.e. principal component analysis. In the alpine lakes application (Ch. 7) CA was also employed to cluster variables.

Using hierarchic agglomerative algorithms, results can be summarized in a *dendrogram*, which provides an easy-to-understand graphical representation of determined groups. An example of dendrogram is shown in **Fig. 2**, where sampling locations coming from a mine waste data set (Ch. 6) where clustered to obtain more homogeneous sub-clusters. Samples name are displayed along the x-axis, while the distance between clusters is displayed along the y-axis.

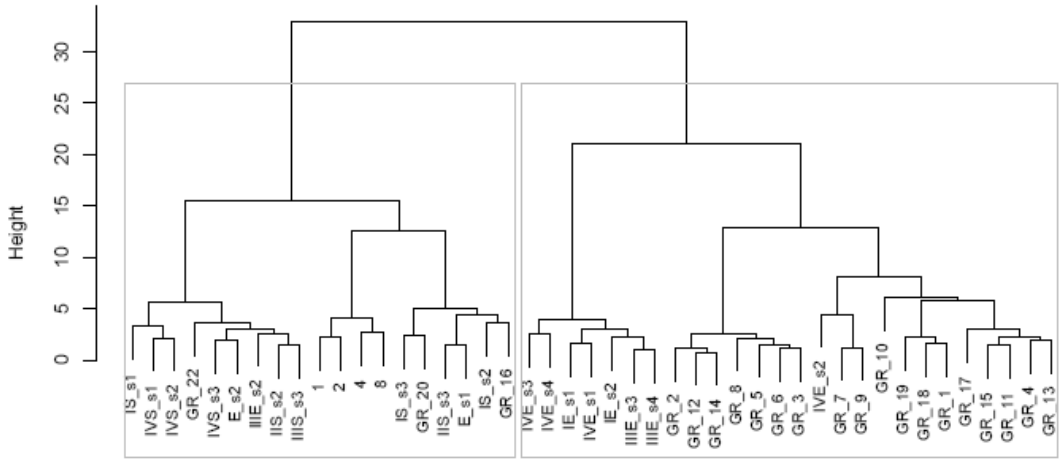


Fig. 2: dendrogram for cluster analysis of the investigate samples in Gromo mine area (Ch. 6)

The main disadvantage of CA is that using different procedures (algorithms and distance measures) on the same data set, can yield to different grouping (Templ *et al*., 2008). However,

CA is a relatively simple technique which provides, in the case of hierarchical agglomerative algorithms, an easy-to-interpret summary of results (dendrogram).

2.2. Distance measures

A distance measure establishes the procedure to quantify the dissimilarity between objects (variables or samples). There are various distance methods to express dissimilarity; the most common used are here below described:

- *Euclidean* distance. It is the most commonly chosen type of distance used in environmental samples analysis. It simply is the square root of the sum of the squared differences in the variables' values;
- *Manhattan* distance. It is also called *City-block* distance and uses the sum of the variable's absolute differences;
- *Chebychev* distance. It is defined by the maximum of the absolute difference in the clustering variable's values.

2.3. Clustering methods

Clustering algorithms can be classified by their searching strategies. A practical distinction is the difference between hierarchical algorithms and partitioning algorithms:

- in *partitioning algorithm*, the number of resulting clusters is pre-defined. The most used partitioning algorithm is the *k*-means, which minimize the average squared distance between the observations and their cluster centres or centroids;
- *hierarchical algorithm* uses a distance matrix as clustering criteria. This method does not require the number of clusters as input information. When groups are formed from the bottom (i.e. the method start with each observation forming a cluster), then the classifications are called **agglomerative**. When the classification starts with the whole data set contained in one cluster, which is then divided into two and more groups, the algorithm is called **divisive**.

The term "Cluster analysis" is often used for the hierarchical agglomerative methods only. Usually these algorithms are preferred in practice. A further classification of them is done in the following sub-section.

2.3.1. Hierarchical agglomerative algorithms

Using a hierarchical classification, results are usually displayed in a dendrogram. A distance matrix to determine similarity between clusters has to be defined. At the beginning, the groups are formed 'from the bottom' where each object represent it own cluster. Then, clusters with the closest distance are joined to form one cluster. Distance between the new groups is computed again to create other clusters. The joining is repeated until one final cluster is formed. A number of different methods are available for linking two clusters. Best known are:

- *Single linkage*. The distance between two clusters is determined by the shortest distance between any two members in the two clusters. This algorithm is also called the Nearest Neighbor algorithm. As a consequence of its construction, single linkage tends to build large groups.
- *Complete linkage*: The complete linkage algorithm considers the greatest distances between any two members in the two clusters, as opposed to the single linkage approach. It is also called the Farthest Neighbor algorithm;
- *Average linkage*: The average linkage algorithm (weighted or unweighted) is a compromise between the two preceding algorithms. The distance between two clusters is determined by the average distance between all pairs of members in the two clusters;
- *Ward*. This method is different from all other methods, in that the distance between clusters is evaluated using an analysis of variance approach. This method attempts to minimize the sum of squares of any two clusters that can be formed at each step.

2.4. Normalization procedures

Among the types of transformation used to obtain a normal distribution, the commonly applied are logarithmic, square root and Box-Cox:

- logarithmic transformation uses natural logarithms of data values to transforms the data set. Generally, the modified log(x+1) transformation is applied, in order to prevent the occurrence of negative results for values less than 1;
- square root transformation consists of tacking the square root of data values. This transformation is generally used when the variable is a count;
- Box-Cox procedure is a power transformation type. It is defined by the following function which varies respect to the parameter λ:

$$y_i^{(\pi)} = \begin{cases} \dfrac{(y_i^{(\pi)} - 1)}{\lambda} & \text{when } \lambda \neq 0 \\ \log(y_i) & \text{when } \lambda = 0 \end{cases}$$

The choice of the best value for λ is generally based on maximum likelihood estimation. Usually, sample skewness is computed to assess whether the data set fit a normal distribution, having skewness in the range of -0.8 to 0.8.

2.5. Standardization procedures

Standardization of geochemical data sets is useful when data-sensitive statistical techniques have to be applied. Some multivariate modelling techniques are in fact strongly dependent on the variable which shows the largest difference in scaling. The most popular procedures are:

- z-scaling, also called autoscaling, computes new data with zero mean and unit variance, according to the equation:

  $$z_i = \frac{x_i - \mu}{s}$$

  where z_i is the standard score of each variable, x_i is the value of variable *i*, μ is the mean, and *s* define the standard deviation. The standardization procedure gives each variable equal weight in the multivariate statistical analysis.
- Pareto scaling uses, differently from z-scaling, the square root of standard deviation as the scaling factor on mean-centred data. With its application, data does not become dimensionless.

Chapter 3

Principal component analysis

3.1. Introduction

Principal component analysis is a multivariate data reduction technique. Its main objective is to reduce the dimensionality of a complex data set, with little loss of information.

Principal component analysis is one of the most commonly used methods for data analyses in environmental sciences. It has been applied in air quality studies (Yu *et al.*, 2000; Motelay-Massei *et al.*, 2003; Pires *et al.*, 2008; Chang *et al.*, 2009) as well as in soil and sediment compartments (Critto *et al.*, 2003; Loska *et* Wiechuła, 2003; Dos Santos *et al.*, 2004; Officer *et al.*, 2004; Bhuiyan *et al*., 2010).

The goal of this technique is to project the original variables in a new reference frame, which make maximum the variance. The new variables, called *principal components* (PCs), are extracted in decreasing order of importance. In this way, the PC with higher variance is projected in the first axis, the second PC on the second axis and so forth. The new variables, which are uncorrelated, represent thus a particular linear combination of the original variables (Davis, 2002).

PCA analysis could be carried out on R or Q-mode. In R-mode analysis, the association among variables is address, while Q-mode analysis focuses on the relationship between observations.

Two methodologies could be implemented to estimate principal components: eigenvalue decomposition or singular value decomposition; further details of their application are given in the following sections.

Principal component analysis is a data-sensitive technique; pre-treatment of data if often necessary to obtain a data set more suitable for its application (Reimann *et al*., 2002). The negative aspect of data pre-treatment is that different transformations can influence PCA results and data interpretation (Reid and Spencer, 2009).

Like in cluster analysis application, a data structure composed by variables with different numerical ranges may produce incorrect PCs, because the variable with the largest variance will have a major influence on results (Reimann *et al*., 2002). Appropriate standardization and/or normalization procedures have to be applied prior to the analysis. In particular, normalization

procedures are used to normalize data distributions, which are often apart to be normal dealing with geochemical data. These procedures are described in § 2.4 and § 2.5.

Outliers should be removed prior to principal component analysis. Even if they can contain important information, they can negatively influence the results of the analysis (Reimann *et al*., 2002).

Sometimes other classification techniques, like cluster analysis, should be used prior to PCA application in order to find more homogeneous sup-population of the original data set (e.g. sub-population determination in the Gromo mine site application, Ch. 6).

3.2. Algorithm

Given a data set with *n* variables, the aim of principal component analysis is to identify as many *n* new variables, called *principal components* (PCs), which are a linear combination of the original variables. The objective is then to reduce the dimensionality of the data set by considering only the first meaningful PCs.

3.2.1. Eigenvector decomposition

Given a generic square matrix **A**, eigenvalues and eigenvectors are a scalar (λ) and a non-zero vector (***v***) so that they satisfy the so-called eigenvalue equation:

$$\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v} \qquad \textbf{Eq. 1)}$$

Let be **X** the data set with *n* variables (e.g. chemicals measurements) and *m* samples. Given a linear transformation **P**, a change of basis could be expressed by the following equation:

$$\mathbf{PX=Y} \qquad \textbf{Eq. 2)}$$

The eigenvector decomposition (EVD) is based on the computation of the covariance matrix, expressed by the following equation:

$$\mathbf{C_X} = \frac{1}{n}\mathbf{X}\mathbf{X}^T \qquad \textbf{Eq. 3)}$$

The covariance matrix elements measure the covariance between all possible pairs of measurements. It is a square symmetric matrix, where the diagonal terms are the variance of particular measurement types and the off-diagonal terms the covariance between measurement types.

The new reference system, expressed by the extracted PCs, could be identified by the matrix **Y** (change of basis). The covariance matrix $\mathbf{C_Y}$ for the new reference system can be computed

similar to the **X** case (**Eq. 3**). Since the objective of PCA resolution is to maximize the variance of PCs, with uncorrelated PCs, all off-diagonal terms in $\mathbf{C_Y}$ should be zero ($\mathbf{C_Y}$ must be a diagonal matrix). To diagonalize **Cy**, PCA assumes that all basis vectors are orthonormal (**P** is an orthonormal matrix).

In this way the problem summarize in the determination of an orthonormal matrix, **P** (**Eq. 1**), such that $\mathbf{C_Y}$ is diagonal. In other words, rewriting $\mathbf{C_Y}$:

$$\mathbf{C_Y} = \frac{1}{n}\mathbf{YY}^T = \frac{1}{n}(\mathbf{PX})(\mathbf{PX})^T = \mathbf{P}(\frac{1}{n}\ \mathbf{XX}^T)\mathbf{P}^T$$

$$= \mathbf{PC_XP}^T \qquad \textbf{Eq. 4)}$$

$$\mathbf{C_YP} = \mathbf{C_XP}$$

The goal of PCA becomes to determine the eigenvectors and eigenvalues of the **X** covariance matrix. In this way the principal components of the **X** matrix are defined by **P** (eigenvectors of $\mathbf{C_X}$) and the diagonal elements of $\mathbf{C_Y}$ matrix (eigenvalues of $\mathbf{C_X}$) correspond to the variance explained by each principal component.

Usually, prior to performing PCA analysis, it is typical to standardize all the variables to zero mean and unit standard deviation in order to eliminate the influence of different measurement scales. This is equivalent to performing a PCA on the basis of the correlation matrix of the original data, rather than the covariance matrix.

3.2.2. Singular value decomposition

Compared with EVD, singular value decomposition (SVD) is a more robust and precise method. Singular value decomposition is generally the preferred method for numerical accuracy and stability (Unonius and Paatero, 1990).

SVD is a matrix factorization technique for decomposing a generic *n* x *m* matrix **A** into three matrices as follows:

$$\mathbf{A}_{mxn} = \mathbf{U}_{mxm}\ \mathbf{S}_{mxn}\ \mathbf{V}^T_{nxn} \qquad \textbf{Eq. 5)}$$

where **U** and **V** are orthonormal matrices ($\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$) and **S** is a diagonal non-square matrix.

It is closely related to PCA being **Eq. 5** similar to **Eq. 4**. The main difference is that in the SVD approach the **X** matrix of **Eq. 4** can be rectangular and the following equation can be solved. $\mathbf{A}\boldsymbol{v} = \sigma\boldsymbol{u}$

where σ are called *singular values* (in a square matrix they equal eigenvalues), and ***u*** and ***v*** are called *singular vectors* (they correspond to eigenvectors in a square matrix).

To relate PCA with SVD starting from the original data matrix **X** we define a new matrix **Y** given by:

$$\mathbf{Y} = \frac{1}{\sqrt{n}}\mathbf{X}^T$$

In this way, by constructing $\mathbf{YY}^T$ we obtain the covariance matrix of **X**. From eigenvector decomposition, we know that the principal components of **X** are the eigenvector of $\mathbf{C_X}$ and hence, computing the SVD of **Y** we obtain:

$$\mathbf{Y} = \mathbf{U\Sigma V}^T$$

and multiplying by the transpose matrix $\mathbf{Y}^T$ (being $\mathbf{VV}^T=\mathbf{I}$) we obtain

$$\mathbf{YY}^T = (\mathbf{U\Sigma V}^T)\,(\mathbf{U\Sigma V}^T)^T = \mathbf{U\Sigma V}^T\,\mathbf{V\Sigma}^T\mathbf{U}^T = \mathbf{U\Lambda U}^T$$

With $\mathbf{\Lambda}=\mathbf{\Sigma\Sigma}^T$. The columns of matrix **U** contain the eigenvectors of $\mathbf{Y}^T\mathbf{Y} = \mathbf{C_X}$. Therefore the columns of **U** are the principal components of **X**.

Like the previous algorithm, selecting only the more important components (those with higher eigenvalues), say the first *h*, data are projected from *m* to *h* dimensions.

3.3. Estimating the number of PCs

Three principal methods are usually used to select the appropriate number of principal components. The first two methods are based on the *scree plot*, the plot of eigenvalues against the corresponding PC (**Fig. 3**). It illustrates the rate of change in the magnitude of the eigenvalues for the PC. The methods used to estimate the number of PCs are here below described:

1. examining the scree plot, the curve tends to decrease fast for the first PCs until it reaches an “elbow”. The number of components to select is given by the PC number at the elbow point;
2. from the scree plot, only PCs with eigenvalue (variance) greater that 1 are retained. This method is usually called *Kaiser criterion*;
3. the last method is based on the cumulative variance. In **Tab. 1** a summary of PCA analysis is given. Since the first few PCs accounts for a large proportion of the total variability, only PCs which represent 80-90% of cumulative proportion of variance are selected.

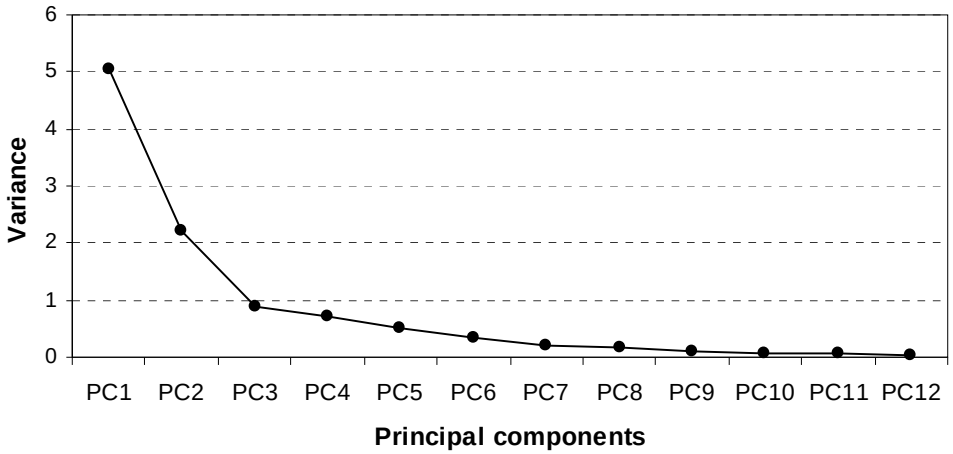


Fig. 3: Example of scree plot for PCA. Eigenvalues, representing the variance, are plotted in the y-axis

Tab. 1: Derived principal components, standard deviation, proportion of variance and cumulative contribution of variance for PCA analysis.

Principal components	*Standard deviation*	*Proportion of variance*	*Cumulative proportion*
PC1	2.25	48.6%	49%
PC2	1.49	21.4%	70%
PC3	0.95	8.6%	79%
PC4	0.85	7.0%	86%
PC5	0.71	4.8%	90%
PC6	0.58	3.3%	94%
PC7	0.45	1.9%	96%
PC8	0.40	1.6%	97%
PC9	0.34	1.1%	98%
PC10	0.28	0.7%	99%
PC11	0.25	0.6%	100%

3.4. Data interpretation

The interpretation of principal components is usually carried out graphically, by means of the *loading plot*. Loadings, which are vectors of the eigenvector matrix, are plotted against each other in order to determine the contribution of each variable in the examined PCs. In fact, the eigenvector or loading matrix contains the cosines of the angle between the original variables and the PCs.

In many statistic software package eigenvectors are converted to correlation coefficient between PCs and the original variables; however the output matrix is called 'loading', which may be eigenvectors or correlation coefficients.

High correlation between PC1 and a variable indicates that the variable is associated with the direction of the maximum amount of variation in the data set. More that one variable might have a high correlation with PC1, explaining its origin (pollution or natural source, chemical process, and so forth). If a variable does not correlate to any PC, this usually suggests that the variable has little or no contribution to the variation in the data set. Therefore, PCA may often indicate which variables are important and which ones may be of little consequence.

The interpretation of PCA results may be subjective. In fact, determined correlation coefficients, or loadings, could be significant for some researcher but not for other.

The main drawback of PCA is the possibility to obtain negative scores, which may not always have a direct physical interpretation (Tauler *et al*., 2004). In fact, factor scores identify the contribution of each sample to the PCs and negative values cannot be interpreted (e.g. if PCs correspond to sources or chemical processes, negative values act as sink).

3.5. Rotations

In PCA a generic rotation is a linear transformation of the original measurements. A rotation was already defined in the factorization problem by means of the **P** transformation in EVD (**Eq. 2**), and **U** and $\mathbf{V}^T$ matrices in SVD decomposition (**Eq. 5**). In these equations, the objective of the rotation was to find the transformation that maximizes the variance of the new variables (PCs). This condition was gained with the diagonalization of the $\mathbf{C}_Y$ matrix (**Eq. 4**). However, in this section we deal with rotations applied only to the subspace defined by the first principal components extracted from PCA analysis.

In fact, rotations are commonly applied after PCA application in order to obtain a clearer pattern of loadings. Typical rotational strategies are *varimax*, *quartimax*, and *equamax*.

The most known analytical algorithm to rotate the loadings is the varimax rotation method proposed by Kaiser (1985). In this case, the objective is to find a rotation that maximizes the variance of the first PCs extracted.

However, the use of rotation after PCA application is questionable. A number of drawbacks were outlined in Jolliffe (2002) and Preacher and MacCallum (2003):

- a rotation criterion must be defined and usually the choice of the Varimax method is due to the default criteria in statistical software packages. Different rotations may produce different results;
- using rotations, the total variance within the rotated subspace determined by the first PCs remain unchanged. With or without rotations, principal components are anyway determined aiming at the maximum variance. Variance is only distribute in a different way after rotations, but in this way, the information carried out by dominant components may be lost;
- results obtained after rotation depend on the number of first PCs forming the subspace;
- the choice of normalization constraint usually applied on the examine data changes the properties of the rotated loadings.

Chapter 4

Positive matrix factorization

4.1. Introduction

Positive matrix factorization (PMF) is a recent approach to multivariate receptor modelling, developed by Paatero and colleagues in the mid-1990s (Paatero and Tapper, 1994; Anttila *et al*., 1995). It has been widely used in air quality studies (Anttila *et al.*, 1995; Polissar *et al.*, 1999; Lee *et al.*, 1999; Xie and Berkowitz, 2006; Begum *et al.*, 2004; Viana *et al*. 2008). In recent years, PMF has also been successfully applied to different geochemical research areas like sediments (Bzdusek *et al*., 2006) as well as soil and water compartment (Reinikainen *et al*., 2001; Vaccaro *et al*., 2007; Lu *et al*., 2008). However, its applications in the last fields is still very poor.

The aim of PMF application is to determine the number of factors (sources or chemical/physical processes) that better explain the input data set variability and to find correlation among the measured variables. Markers for pollution sources as well as hidden information of the data structure may also be identified.

One of the most important characteristics of positive matrix factorization is the use of the uncertainties matrix which allows individual weights for all the input variables to solve the factorization problem (Paatero and Tapper, 1994). This becomes increasingly important with the introduction of the Guide for Expression of Measurements (GUM) and the derived Guide for Quantification of Analytical Measurements (QUAM), which are nowadays commonly accepted references underlying numerous national and international standards (ISO/IEC, 2008; Ellison *et al.*, 2000).

In contrast to CA and PCA, the use of data uncertainties makes PMF a non-data-sensitive technique where non representative data, such as below-detection limit, missing values and outliers, could be managed by the model reducing their importance (Paatero and Tapper, 1994), and data characterized by skewed distribution could be appropriately weighted rather than normalized (Huang and Conte, 2009).

Moreover, the mathematical algorithm of PMF prevents the occurrence of negative factor loadings and scores, which can arise from PCA analysis, allowing more physically realistic solutions (i.e. positive factor profiles) (Reff *et al*., 2007).

Different approaches to resolve the PMF model have been studied: 2-way, 3-way and N-way algorithms. The firsts programs developed by Paatero, solving the 2-way and 3-way problems, are called *PMF2* and *PMF3*, respectively (Paatero, 1997; Paatero, 2004a; Paatero 2004b). Later on the algorithm has been extender to arbitrary multilinear models with the *Multilinear Engine* (ME) program (Paatero, 1999). In the latest years, new custom algorithms were developed by other starting from Paatero's PMF resolution (e.g. Bzdusek *et al.*, 2006). Moreover, given the importance of receptor models in scientific research, the United States Environmental Protection Agency (US-EPA) developed a standalone version of PMF, EPA PMF 3.0, for the resolution of 2-way problems. It was conceived for atmospheric studies and it is freely distributed (Norris *et al.*, 2008). EPA PMF 3.0 is based on ME-2 (ME second version; Paatero, 2007c)

4.2. PMF model

The principle of PMF algorithm start from the basic mass balance equation which, in a two-way problem, given an input *n*x*m* data matrix **X**, is described by the following equation:

X = **GF** + **E**

or, in component form:

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \qquad i = 1\ldots m;\ j = 1\ldots n;\ k = 1\ldots p \qquad \textbf{Eq. 6}$$

where g_{ik} and f_{kj} are the elements of the so-called factor scores and factor loadings matrices, respectively; e_{ij} are the residuals (i.e. the difference between input data and predicted values) and p is the number of resolved factors (Paatero, 1997; Paatero, 2007a). Usually, in environmental studies, the **X** matrix corresponds to known m chemical measurements over n time periods or n sampling locations, **G** represent the p sources' contribution and **F** is a matrix containing source profiles for the p sources and m chemical variables. As stated in **Ch. 1** no priori information about **F** and **G** matrices is required by the model.

PMF solves **Eq. 6** via a weighted least squared algorithm. It iteratively computes **G** and **F** that minimize the so-called *object function Q*, defined in Paatero (1997) and given by the (simplified) equation:

$$Q(E) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{e_{ij}}{\sigma_{ij}} \right)^2$$

where σ_{ij} is the error estimate (uncertainty) associated with each data. The scaling of data using individual error estimates optimizes the information content of the data by weighting variables by their importance. In this way, problematic data could be opportunely weighted.

Additionally, all **G** and **F** elements are constrained to be positive allowing positive source profiles and source contributions in order to make physically realistic the solution (e.g. sources may not emit negative amounts of chemical substances; Paatero and Tapper, 1994).

In this way the PMF problem is identified by the minimization of Q(E) with respect to **G** and **F**, and under the constraint that all their elements must be non-negative.

4.2.1. Resolving algorithm

The PMF2 program was base on alternating regression (AR) algorithms. In AR, starting from pseudo-random initial values, one of the factor matrices, say **G**, would be held constant, while the Q object function is being minimize respect **F**. Then **F** would be held constant while **G** is iteratively estimated. This process continues until convergence (Paatero and Tapper, 1993). In order to reduce the time required for computation, Paatero and Tapper improved the performance of AR algorithm introducing a third step where both **G** and **F** changes simultaneously. Considering **ΔG** and **ΔF** two arbitrary matrices in the factor space of **G** and **F**, the algorithm perform the minimization of Q(**G**+**ΔG**, **F**+**ΔF**) allowing **ΔG** and **ΔF** to change simultaneously.

Since the convergence of the AR solution can be very slow, the PMF2 algorithm was created by Paatero and collegues as a generalization of the AR algorithm. PMF2 is able to simultaneous vary the elements of G and F in each iterative steps and have a faster convergence. Here, the Q object function assumes a more complicated formula with the inclusion of four additional terms: two for the implementation of the non-negativity constraint of **G** and **F**; and two to reduce the rotational ambiguity (see rotations, § 4.2.2).

A brief explanation of PMF2 method is given, but for a detailed description refers to Paatero, 1997. The new object function, called *enhanced object function* is defined as:

$$\begin{aligned}\overline{Q}(E,G,F) &= Q(E) + P(G) + P(F) + R(G) + R(F) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 - \alpha\sum_{i=1}^{m}\sum_{k=1}^{p}\log g_{ik} - \beta\sum_{k=1}^{p}\sum_{j=1}^{n}\log f_{kj} \\ &+ \gamma\sum_{i=1}^{m}\sum_{k=1}^{p} g_{ik}^2 + \delta\sum_{k=1}^{p}\sum_{j=1}^{n} f_{kj}^2\end{aligned} \qquad (3.2.4)$$

where P(**G**) and P(**F**) are called *penalty functions* and prevent the elements of the factor matrices **G** and **F** from becoming negative. R(**G**) and R(**F**), called *regularization functions*, are used to

remove some rotational indeterminacy and to control the scaling of the factors. The α, β, γ and δ coefficients control the strength of their respective functions. For efficiency reasons the log function of the penalty term was approximated by a Taylor series expansion up to quadratic terms (Paatero, 1997).

During each iteration step, Paatero chose to use the Gauss-Newton and Newton-Raphson numerical methods and the Cholesky decomposition. Between steps, rotational sub-steps are performed: a rotation (a linear transformation in PMF jargon; Paatero and Tapper, 1993) **T** and its inverse $\mathbf{T}^{-1}$ can be applied to the factor matrices so that the **GT** and $\mathbf{T}^{-1}\mathbf{F}$ minimize the enhanced object function. In this way, the residual of the fit do not change and rotations increase the speed of computation.

4.2.2. Rotational ambiguity

Despite the non-negativity constraint of **G** and **F** elements, PMF solutions may not be unique but is affected by rotational ambiguity.

Given a linear transformation (rotation) **T**, the expression $\mathbf{GF} = \mathbf{GTT}^{-1}\mathbf{F}$ represent a pair of factors, **GT** and $\mathbf{T}^{-1}\mathbf{F}$, which are are '*equally good*' (same goodness of fit) as the original pair, **G** and **F**. Actually there are different possible rotations so the objective is to determine the optimal solution that better represents the problem under analysis. A given $t_{ij}>0$ (positive **T** matrix element) creates rotations imposing additions among loadings (**F** rows) and subtractions among the corresponding scores (**G** columns); when $t_{ij}<0$ the role of the matrices is exchanged (Paatero *et al.*, 2002).

An infinite number of rotations may exist satisfying the non-negativity constraint.

In PMF2 algorithm rotations are implemented during iterative steps by means of the so-called *FPEAK* parameter, which can assume positive or negative value (the zero-value correspond to the un-rotated solution, called *central solution*).

4.3. Error estimates

PMF is a weighted least square model with the property to use individual error estimates to weight data points.

PMF2 program allows to directly introducing the error estimates matrix, which can be either previously determined by the user or computed by setting different parameters in the PMF2 initialization file (.INI file). In the last case, the combination of three different numerical codes,

called C1, C2 and C3, defines the so-called *Error Model*s (*EM*s), which determine different formulas used to compute the error estimates matrix. The C1, C2 and C3 codes (see **App. B** for their identification into the .INI file) are associated to **T**, **U** and **V** arrays, respectively, which are defined by the user.

In the simplest case in which all the input data have the same uncertainty, only the one-value C1, C2 and C3 codes value have to be set. Alternatively, if individual uncertainties are evaluated the corresponding **T**, **U** and **V** matrices are used. The values C1 and t_{ij} are expressed in same units of x_{ij} (input data), while C2 and C3 and the arrays U and V are dimensionless. Usually, the **V** array contains relative errors of data point and **U** (or C2 value) is used only in rare cases.

Depending on the used EMs, the error estimates matrix (**S**) could be computed either before the algorithm computation (EM= –12) or during each iterative steps, using fitted values in place of the input data (EM = –10, –11, –13, –14). Following, a description of the available error models:

- *EM = –12*. The equation used to determine the error estimates matrix elements is given by:

  $$s_{ij} = t_{ij} + u_{ij}\sqrt{|x_{ij}|} + v_{ij}|x_{ij}|$$

  The **T** matrix corresponds to the x_{ij} analytical uncertainties matrix, while **V** contains relative errors.

- *EM = –10*. This structure is used when it is assumed that data and uncertainties have a lognormal distribution. The **S** matrix is iteratively calculated by:

  $$s_{ij} = \sqrt{t_{ij}^2 + 0.5v_{ij}^2|y_{ij}|\left(|y_{ij}| + |x_{ij}|\right)}$$

  T represents typical measurement errors, while **V** contains the geometric standard deviation logarithm. During the iterative steps y_{ij} is the fitted values.

- *EM = –11*. The following formulation is used when the date set fit a Poisson distribution. Being μ= **GF**, the error matrix **S** is computed by:

  $$s_{ij} = \sqrt{\max\left(|\mu_{ij}|, 0.1\right)}$$

- *EM = –13*. The error matrix is computed using the same equation of EM = –12. The difference being that in the EM = –13 structure the error estimates are computed iteratively, replacing the x_{ij} input data with the y_{ij} fitted values.

- *EM = –14*. The following equation was use to determine **S** matrix:

  $$s_{ij} = t_{ij} + u_{ij}\sqrt{\max\left(|x_{ij}|, |y_{ij}|\right)} + v_{ij}\max\left(|x_{ij}|, |y_{ij}|\right)$$

This option is recommended in environmental work as an alternative method to the *EM = –12*, although the processing time is greater.

When the error estimate matrix is read from an external file (i.e. the matrix is computed by the user using literature methods) only the **T** array is read, setting C2 = C3 = 0 and EM = –12.

4.4. Non-representative data

4.4.1. Below detection limit and missing data

Typically, environmental data sets can contain BDL and/or missing values. To make use of their information content, opportune estimates for their values and uncertainties must be determined.

Usually, when '<DL' values are present within a data set, use of uncensored data (if available) may be preferred (Farnham *et al*., 2002); otherwise proper data estimates are employed. Different types of data and uncertainty estimates can be found in literature; some examples are given in **Tab. 2** and **Tab. 3**. It can be observed that data estimated are the same for all given the examples; in fact DL/2 is a very common choice to substitute BDL data.

Detection limit is a common quantity used for computing the uncertainty matrix; in the examples given in **Tab. 3**, it specified the error estimates for low data value.

A combination with literature formulas and EMs could be determined, providing good BDL and missing data uncertainty estimates in **T** and **V** matrices.

Moreover, PMF2 program allows an automatic handling of missing value and BDL by the use of the optional parameters *Missingneg r* and *BDLneg r1 r2*, respectively. For detailed information see Paatero, 2004a. However these options must be used with caution.

Tab. 2: examples of non-representative data estimates. x_{ij} are the input measurements, DL is the method detection limit and $\overline{x}_{ij}$ is the geometric mean of measurement.

	Determined Values	*BDL data*	*Missing values*
Polissar *et al*. (1998)	x_{ij}	$DL_{ij}/2$	$\overline{x}_{ij}$
Xie and Berkowitz (2006)	x_{ij}	$DL_{ij}/2$	$\overline{x}_{ij}$
Polissar *et al*. (2001)	x_{ij}	$DL_{ij}/2$	$\overline{x}_{ij}$

Tab. 3: an example of uncertainties estimates. u_{ij} are analytical uncertainties, DL is the method detection limit and $\overline{x}_{ij}$ is the geometric mean of measurement. C2 is a percentage parameter, while a and b are scaling factors, both determined by trial and error.

	Determined values	*BDL data*	*Missing values*
Polissar *et al*. (1998)	$DL_{ij}/3 + u_{ij}$	$\overline{DL}_{ij}/2 + DL_{ij}/3$	$4 \cdot \overline{x}_{ij}$
Xie and Berkowitz (2006)	$DL_{ij}/3 + C2 \cdot x_{ij}$	$\overline{DL}_{ij}/2 + DL_{ij}/3$	$4 \cdot \overline{x}_{ij}$
Polissar *et al*. (2001)	$\sqrt{a_j u_{ij}^2 + b_j DL_{ij}^2}$	$b_j DL_{ij}$	$25 \cdot \overline{x}_{ij}$

4.4.2. Outliers

Outliers are extreme values that differ from the mean trend of all the data. They can occur for various reasons and can be 'true', in the case of a contamination or pollutant source (i.e. mineralization) or 'false', if resulting from sampling or analytical error. In either case, they can have a significant influence on multivariate analysis results.

To overcome this drawback, PMF offer the so-called *robust mod*e which act reducing the outliers influence. In this case, outliers are dynamically reweighted during the iteration by means of the *Huber influence function*, which modify the Q formulation (Paatero, 1997). The Hubert function limits the maximum strength that each data can bring to the fit and is defined by:

$$\psi^H(r_{ij}) = \begin{cases} -\alpha & \text{if } r_{ij} < -\alpha \\ r_{ij} & \text{if } -\alpha \le r_{ij} \le \alpha \\ +\alpha & \text{if } r_{ij} > \alpha \end{cases}$$

where α is the outlier distance (the distance for classifying the observation as outliers) and $r_{ij} = e_{ij}/\sigma_{ij}$ are the scaled residues. The object function corresponding to ψ^H is denoted by Q^H and the least square formulation becomes:

$$Q^H(E) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{e_{ij}}{h_{ij}\sigma_{ij}}\right)^2 \qquad h_{ij}^2 = \begin{cases} 1 & \text{if} \left|e_{ij}/\sigma_{ij}\right| \le \alpha \\ \left|e_{ij}/\sigma_{ij}\right|/\alpha & \text{otherwise} \end{cases}$$

In this way, outliers are handled as they stay at the distance $\alpha\sigma_{ij}$ from the fitted value. This method however is not applied to negative outlier (data showing very low values respect the mean observations).

4.4.3. High noise variables

In environmental studies it may happens either that some variables present a higher noise than others or the noise is greater than the signal.

In Paatero and Hopke (2003) the signal to noise ratio (S/N) was used to classifies variables: *weak* variables contain signal and noise in similar quantities; *bad* variables contains much more noise than signal. In numerical terms weak variable have 0.2<S/N<2 and bad variables S/N<0.2. If detection limits are known the S/N ratio could be computed by means of the following equation:

$$\frac{S}{N} = \frac{\sum_{\{i|x_{ij}>\delta_{ij}|\}} x_{ij}}{\delta_j n_{DLj}}$$

where, in the *j* column, n_{DLj} is the number of below-detection-limit data and δ_j is the mean detection limit.

Paatero and Hopke (2003) recommended to downweight weak variable by a 2 or 3 factor. Bad variable could be omitted from the analysis or must be downweighted by a factor between 5 and 10.

4.5. Explained variations

Explained Variation (EV) is a dimensionless quantity which describes the relative contribution of each factor in explaining a row (EV of G matrix) or a column (EV of F matrix) of the input data set, X. On the other hand, residuals could be considered to form a fictitious *(p+1)* factor called 'not explained variation' (NEV) and representing the unexplained part of the data set by the *p*-factor model.

The EV values range between 0 and 1 corresponding to no explanation and complete explanation, respectively. The explained variation matrices are defined in Paatero (2004b). In the **G** matrix case, EVG and NEVG are given by the equations:

$$EVG_{ik} = \frac{\sum_{j=1}^{m} \left| g_{ik} f_{kj} \right| / s_{ij}}{\sum_{j=1}^{m} \left(\sum_{h=1}^{p} \left| g_{ih} f_{hj} \right| + \left| e_{ij} \right| \right) / s_{ij}} \qquad \text{for } k = 1, \ldots, p$$

$$NEVG_{ik} = \frac{\sum_{j=1}^{m} \left| e_{ij} \right| / s_{ij}}{\sum_{j=1}^{m} \left(\sum_{h=1}^{p} \left| g_{ih} f_{hj} \right| + \left| e_{ij} \right| \right) / s_{ij}} \qquad \text{for } k = p + 1$$

The first equation gives information about the relative contribution of each factor (1, …, p) to the i^{th} row of X; in the case of a environmental data set containing *m* chemical measurements in *n* samples, EVG_{ik} describe the amount of i^{th} sample explained by the k^{th} factor. Opposite, NEVG describes the amount of i^{th} sample not explained by the *p*-factor model. By definition, EVG and NEVG sum up to one.

Similar equations are used to determine EVF and NEVF matrices, where the sum is computed over the *i* index. In the case of environmental data sets, EVFs are a measure of the relative contribution of each variable in the determined sources. They are useful outputs providing a qualitative identification of the sources; a factor explaining a large amount of one or more variables can be identified according to their origin. Moreover, NEVF value was used to identify variables which were not explained by the p-factors model. However, it is a practical rule to consider unexplained a variable when its NEVF value exceeds 0.25.

4.6. Initialization file

PMF2 program runs under DOS environment (it is not an installation program). An initialization file, with *.INI* extension is used to read and process the input matrices and other input parameters. An example of .INI file is given in **App. B**. For more detailed information on *.INI* file compilation refer to Paatero (2004a, 2004b) user's guide. Here a summary of most important parameters is given. The .INI file can be split in three main sections, defined in **App. B**: input parameters, input and output files, and optional information.

4.6.1. Input parameters

In the first part of the .INI file code, dimension of the input data matrix and the number of factors to be computed must be set. Usually different numbers of factors are tested, changing every time the .INI file. The "number of repeats" value is set equal to the number of continuous computations to repeat in every run. According to the *pseudorandom seed* parameters, pseudorandom numbers are generated to initialize the algorithm.

FPEAK parameter defines the rotational degree and must be changed every time a new rotation would be tested. The central solution is achieved with FPEAK=0 (default value).
With the "Mode" parameter set to "T" (true) the PMF computation is carried out in the robust mode, which provide re-weight of possible outliers contained in the input data matrix (§ 4.4.2). An outlier distance can be set to define the outliers threshold; usually α values are set to 0.2, 0.4 (default value) and 0.8. Alternatively, two different thresholds for positive and negative residues, respectively, can be defined by means of the optional parameter *outlimits*; optional parameters are inserted at the end of the .INI file (**App. B**, *optional information*).
In the same section of the .INI file, error model is selected. C1, C2, C3 codes and EM value permits to input different error estimates, either based on existing structures or computed by the user (for more details see § 4.3).
The last information to introduce in the input parameters section of the .INI file is given by the i*teration control table*. This table control the convergence of the model by means of four parameters. Three level of convergence are required, the last one being the more restrictive. For a detailed explanation of the iteration control table refer to Paatero (2004a, 2004b). Usually the default convergence criteria are not modified.

4.6.2. Input and output files

In this section, input file are introduced writing their name and extension. Usually, the *.txt* extension is used. Also formats for both input and output files are defined.
The outputs are organized in *.txt* file according to the chosen format. The most important outputs are **G** and **F** matrices, their explained variations, Q value for each run, *rotmat* matrix and the scaled residual matrix. A *.log* file, which contains possible errors occurred during the computation, is also produced.

4.6.3. Optional information

Factor matrices can be normalized according to different options:

- None: no normalization;
- Max**G** = 1/Max**F** = 1: the maximum absolute value in each **G/F** column is equal to the unity;
- Sum|**G**| = 1/ Sum|**F**| = 1: the sum of the elements absolute value in each **G/F** column is equal to the unity;

- Mean|**G**| = 1/ Mean|**F**| = 1: the mean value of the elements absolute value in each **G**/**F** column is equal to the unity.

With normalization the **GF** product did not change. When dealing with results from different runs, it can happen that produced factors (**G** columns and **F** rows) are displayed in a random order. For better results comparison, in order to show factors in the same position of the output file, the optional commands *sortfactorsg* or *sortfactorsf* are used.

However, it is suggested to not use these commands when examining different rotations changing FPEAK parameter. In this case, better results are obtained starting from the lowest FPEAK and use, as a starting point for the following rotations, the results obtained from the previous computation; this is done by means of the *goodstart* parameter.

4.7. Determination of the optimum solution

In this section the parameters involved in the selection of the optimum solution will be investigated. There are in fact several parameters which pertain to the determination of **G** and **F** matrices and the best way to solve the problem is to investigate the most significant combinations of them.

The first step for the determination of the best fit is the computation of different solution varying the number of factors to be considered. At the beginning central solution (with FPEAK=0) are examined. Usually, from 2 to 8-10 factors are considered. The following step consists in the investigation of the rotational degree, varying the FPEAK parameter, for the more significant solutions.

The combination of all the examined parameters used to select number of factors and rotation consent to draw conclusion about the best PMF fit which better characterize the data set under examination.

4.7.1. Determination of the number of factors

Among the computed central solutions obtained varying the number of factors, only the most significant solutions were retained for further analysis of the rotational degree. In this section, output parameters were examined to help reducing the range of possible solution.

Analysis of Q value

In weighted-least-square problems, if the data uncertainties are properly defined, the Q function should be distributed as a chi-square (χ^2) distribution. In the two-dimensional approach, the free parameters of the **GF** product is given by (n + m)x p. Considering also the rotational ambiguity by means of the introduction of the T matrix (pxp) the number of free parameters become (n+m–p)x p. Given the Q expression, the resulting degrees of freedom are $\upsilon = n$xm – (n + m – p)xp = (n – p)x(m – p) (Paatero and Tapper, 1993) and the expected Q (being a χ^2 value) is given by:

$$Q_{exp} = (n - p)x(m - p)$$

If the data matrix is expected to be very large then $Q_{exp} \approx m$xn, that is the expected Q value could be approximated to the number of data points.

In this way, Q_{exp} value gives important information about the quality of the fit because the optimal solution should have a Q not too different from Q_{exp}. Too high or too low (less than Q_{exp}) Q value indicates that the chosen number of factor is too low or too high, respectively. However, when a dataset contains much weak variables or the uncertainties are not well defined, Q can be not comparable to Q_{exp} (Bzdusek *et al.*, 2006).

To extract information about the number of factors to retain, Q/Q_{exp} is plotted against the number of factors examined, as show in the example given in **Fig. 4**

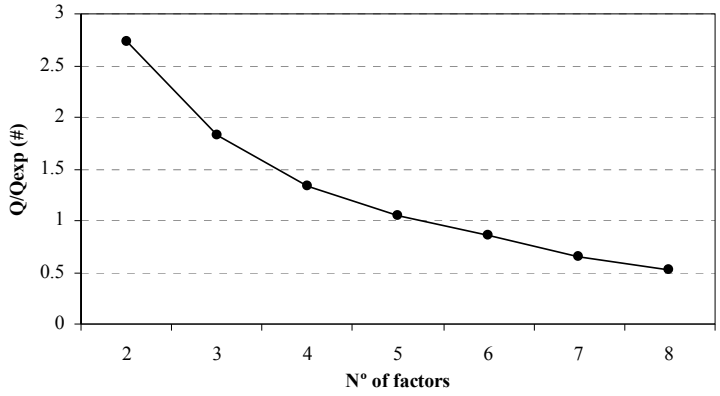


Fig. 4: Q/Q_{exp} for central solution against the number of factors examined

From **Fig. 4** it can be observed that Q/Q_{exp} has a greater slope passing from factor 2 to 3. Moreover, for solution with more than 5 factors resolved the ration is less than 1 suspecting that the chosen number of factor is too high. In this way, we could restrict the range of possible solution from 3 to 5 factors.

In addiction, stability of Q value can be assessed examining the Q variation for each run performed with the same number of factor. Usually from 10 to 15 runs were computed. If local

minima occur, they must be examined. However, local minima are usually correlated with a too high number of factors resolved.

Analysis of scaled residuals

Scaled residuals can be used to detect data anomalies, such as outliers, and to correct too low or too high data uncertainties (Juntto and Paatero, 1994). If data follow a normal distribution and uncertainties are properly determined, the scaled residual frequency plot shows a random distribution with the majority of values located in the range -2, +2 (Juntto e Paatero, 2004).

If their value fluctuate outside this range it is possible that the chosen number of factors is not the best one, that some outliers occur or that uncertainties are set too low for the particular variable. Contrary, if scaled residuals distribution is very narrow, it is possible that uncertainties are too large and it is better to reduce their values. However, narrow distributions can also arise when a variable is explained by a unique factor. This situation may occur both naturally but also when high uncertainties have been specified for a noisy variable (Paatero, 2004a).

However, it is necessary to treat scaled residuals results with caution since it could happen that a bad distribution is due to a natural condition rather than to poor uncertainties (Huang *et al.*, 1999). Referring to the data set analysed in Ch.7 where different Italian lakes sediments were analysed, the residual distribution of Pb variable presented a bimodal character (**Fig. 5**). In this case, the bimodal distribution refers to true outliers which characterize a strong Pb concentration in a particular lake. Actually, bimodal distributions reflect the original spatial distribution (Polissar *et al.*, 1998).

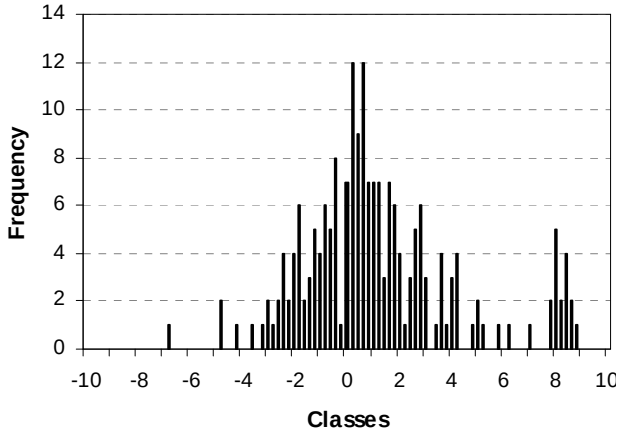


Fig. 5: plot of scaled residual distribution for Pb concentrations measured at different Italian lakes sediments

IM and IS

In order to reduce the range of the meaningful solutions, the *IM* and *IS* parameters are computed using the expression defined in Lee *et al*. (1999). Starting from the scaled residual matrix **R** (r_{ij} elements), IM and IS are given by:

$$IM = \max_{j=1\ldots m}\left(\frac{1}{n}\sum_{i=1}^{n} r_{ij}\right)$$

$$IS = \max_{j=1\ldots m}\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(r_{ij}-\bar{r}_j\right)^2}\right)$$

where $\bar{r}_j$ is the mean over the *i* row.

Examining the IM and IS equations, it can be observed that IM represents the *j* variable with greater scaled residuals mean, while IS reproduces the *j* variable with greater scaled residual standard deviation. In this way, IM define the less accurate fit and IS the more imprecise fit.

Plotting these parameters against the number of factors, solution with high IM and IS values could be rejected (Lee *et al*., 1999). Moreover, IM and IS could show a drastic decrease when the number of factors increase up to a critical value.

Analysing IM and IS values from an example data set, reported in **Fig. 6**, we can observe a rapid decrease of IM from 3 to 4 number of factors and a further decrease from 5 to 6, while IS show a first stationary step between 3 and 4 factors extracted. Combining the results solutions with 3 to 5 number of factors could be further examined.

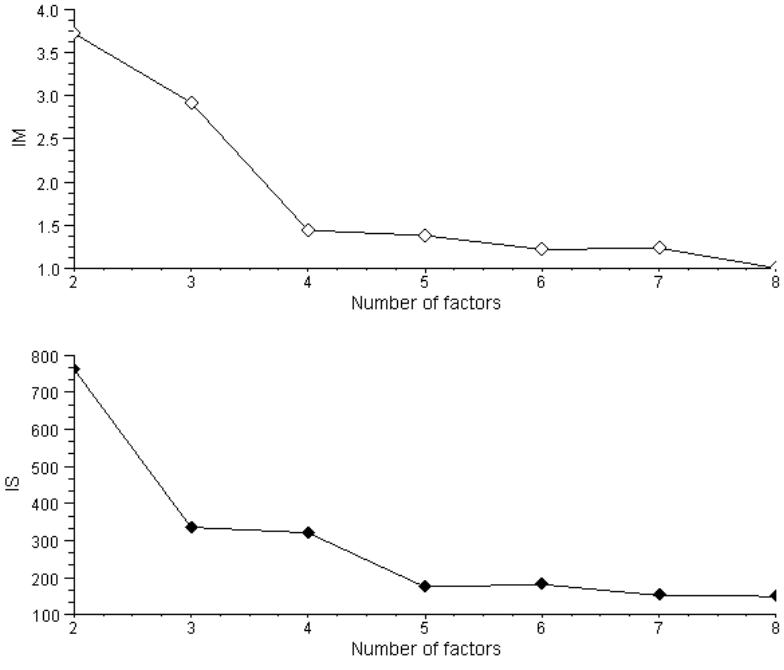


Fig. 6: IM and IS plot vs number of factors

Rotmat

The rotmat matrix indicates the rotational freedom of the solution. Plotting the matrix element with greater value (greater rotational freedom, *MaxRotMat*) for each examined number of factors we gain information about the rotational freedom of the solutions (Lee *et al*., 1999). In this way, it possible to reject solutions that exhibit a rapid change in their rotational degree.

In **Fig. 7** an example of MaxRotMat plot is shown; it can be noticed that solutions with 2 and 8 factors show a rapid positive change in the parameter value. This is compatible with a higher rotational ambiguity and those solutions could be rejected.

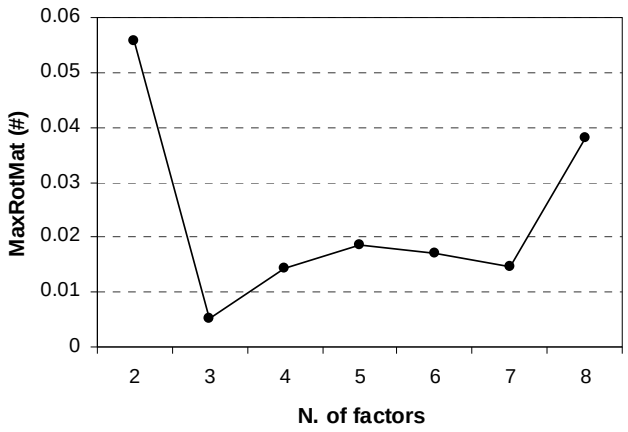


Fig. 7: MaxRotMat value for differtn number of factoer tested by PMF

Not explained variation

Not explained variations represent the portion of data variability not explained by the *p* factor model. When a variable shows high NEVF values, say more than 25-30%, it is not characterize by the model. In this case, a new additional factor could be necessary for a better resolution of the variable, but it could also happens that the variable is not explained because it contains many non-representative data.

4.7.2. Controlling rotations

The rotational degree of PMF solutions can be controlled by means of the FPEAK parameter, which can assume both positive and negative values. Usually, in the majority of PMF applications, rotations are evaluated in the range -1 <FPEAK< +1, with a 0.1 or 0.2 incremental step.

Usually, pseudorandom numbers are used to initialize the PMF2 algorithm. However, when different rotations have to be tested, the use of pseudorandom number is not suggested. Their use

can in fact cause different local minima and the factors to appear with a different index in every rotated solution, making the comparison of rotations more complicated. Paatero *et al*. (2002) suggests the following scheme when operating with rotations:

- perform different initialization runs with pseudorandom value and FPEAK = 0 (central solution) in order to evaluate the Q stability;
- choose the best central solution and use it as a starting point for the data processing with rotations. This is done using the *goodstart* parameter.

Once the range of most meaningful central solution was determined, different rotations can be tested on them. The problem become now to determine the best combination between number of factors and rotation that better characterize the examined data set. A set of parameters is analysed to reject the less appropriate rotations.

Assessing the increase of Q

Q values for rotated solution may show higher values than the central solution (Paatero *et al*., 2002). A customary trend of Q value respect the FPEAK parameter was described by Paatero *et al*. (2002): starting from the central solution Q value initially increases with a little slope up to a certain rotation, at which it start to increase quickly. At the rotations after the change of slope, the factor matrices tend to be distorted because of the non-negativity constraint and the rotations could be rejected. However further experience is needed in order to have a best knowledge in choosing FPEAK values. Anyway, this could be a helpful tool to make a first step decision on the rotate solutions to be considered.

It is not possible to define a precise rule, based on Q value, that allow us to decide when a rotation is to rejected but, as a practical decisional step, we could considered forbidden rotations that show an increase of Q values for more than 10% respect to the central Q (Q_{cen}, Paatero *et al.*, 2002).

In **Fig. 8** an example of Q variation for rotated solutions is given. Even if the ratio Q_{rot}/Q_{cen} gets an increase in the positive FPEAK direction, the difference between rotated and centra Q is lower that 1% and all the rotations can be considered significant.

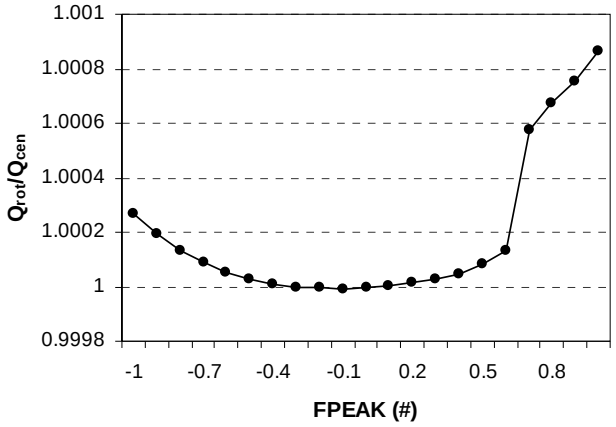


Fig. 8: Q rotational and Q central ration varying the FPEAK parameter

Scaled residual

Similarly to the inspection of the number of factors extracted, scaled residuals can be inspected to check rotations. However, as already explained, some deviation from a normal distribution in the range -2 : +2 may be due to natural data trends.

IM, IS and rotmat

The parameters IM, IS and MaxRotMat, previously described, are used to select the most meaningful range of FPEAK values. The best rotations should have low and stable IM and IS values, representing the more accurate and precise fits, respectively.

In **Fig. 9**, an example is given.

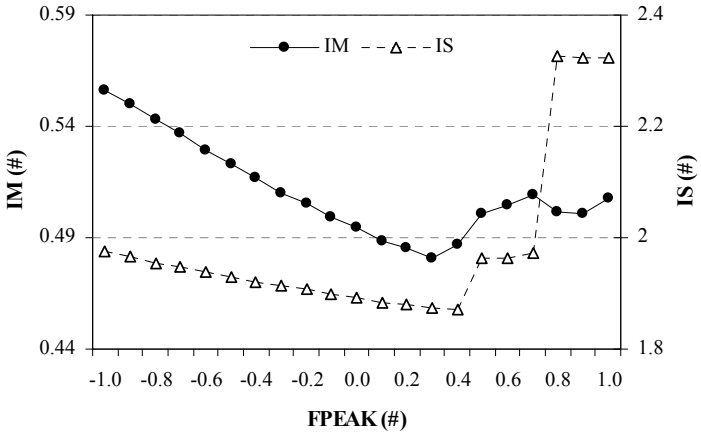


Fig. 9: IM and IS parameters varying FPEAK value

Rotmat matrix is inspected choosing the maximum value for each examined rotation. Plot of MaxRotMat against FPEAK value give information about the rotational ambiguity of solutions. Rotation with lower MaxRotMat values will be favoured (Lee *et al.*, 1999).

G-plots

A graphical approach could be applied on **G** matrix elements in order to select between rotations; this method is called *G space plotting* (Paatero *et al.*, 2005). It is made the assumption that the determined factors are uncorrelated each other. Actually, there is always a weak correlation between pairs of factors, called *weak independence*. The goal of this method is to reject the rotations that give correlation between pair of factors. Scatter plots of **G** matrix elements for two different factors were examined. All the points lie in the positive quadrant because of the non-negative constraint and, if the plotted factors are uncorrelated, the straight lines passing thought the origin of axes and including all the point between them should approximate the Cartesian axes. These lines are called *edges* and scatter plots with edges nearest the axes are those relating to the optimum rotation.

However, there may be physical situations where oblique edges can naturally occur and a good knowledge of the problem under analysis may help in the scatter plot interpretation. Also, edges near the axes do not guarantee that the solution is unique (Paatero *et al.*, 2005).

In **Fig. 10** an example of two G plots. In the graph on the left side the two factors are uncorrelated, with edge; scatter-plot on the right show some correlation between factors.

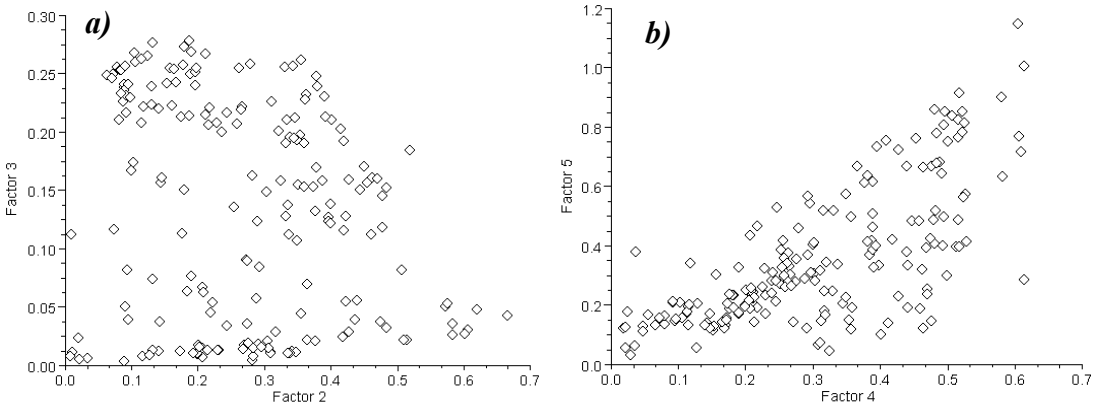


Fig. 10: G plot between two uncorrelated factors (a) and two factors showing correlation (b)

4.7.3. Fkey: a *priori* information

An alternative approach for controlling rotations is the use of a priori information (Paatero *et al.*, 2002). Selection among different solutions given by different FPEAK values may be performed by the knowledge of some information on the problem under analysis (e.g. information obtained from preceding studies).

A priori information may be input within the algorithm through the use of the Fkey matrix that works pulling down to zero some **F** elements. Like this, Fkey matrix guides the analysis towards a more understanding solution/rotation. For example, if it is known that one or more variables have a null contribution on some factors, this information can be implemented through the Fkey matrix in order to force the variable to the known values (Lee *et al.*, 1999). However forcing to zero the elements in the **F** matrix seems to increase the frequency of local minima, giving rise to multiple problem solution (Paatero, 1997).

Lingwall and Christensen (2007) studied the a priori information effects using simulated experiments. The results showed that resolved factors could be improved when the pulling to zero elements is performed on 'clean data' (i.e. data with low uncertainties and not affected by unidentified source). However, a worse the fit could be obtained if the information provided in the Fkey matrix is not correct.

Chapter 5

LIMS

A laboratory information management system (LIMS) is a database system used in laboratories for the management of instruments, individual samples and the information obtained on them with different analytical tools.

In JRC-IES laboratories, where a great number of samples have been collected and tested, one of the main tasks of LIMS is the automated production of barcodes for sample identification. During monitoring campaigns carried out at JRC-IES (e.g. the FATE SEES campaigns described in Ch. 10) a specific protocol was defined to establish the methodology for dispatch of samples from the JRC to other organizations, either for samples collection (dispatch of empty containers) or samples external laboratory analyses. LIMS was successfully used at this stage to register and label empty containers before sampling and to register samples information achieved after samples collection. Furthermore, LIMS integrates with a barcode reader which simplifies the laboratory workflow.

After samples analysis, LIMS is used to accurately keep track of results which, after validation, are archived in the system.

LIMS is also used to register laboratory instruments/equipments and to store and program their maintenance.

5.1. Sample labels

Prior to dispatch or collection of samples, sample labels were created. Labels identify each sample in a unique way by means of the barcode automatically generated in LIMS. An example barcode-label is show in **Fig. 11**.

A new barcode-label must be created whenever pre-treatment procedures are applied to sample sub-sets. Indeed, in this case a new sample with different matrix is created and must be registered differently from the original sample.

Fig. 11: Example of barcode-label created in LIMS software

Referring to **Fig. 11** numbering, the barcode-label contains the following information:

1. Sample ID or bar code: is automatically generated by the system and identify the combination of sample/label. It is unique for each sample/label combination;
2. Name of the project;
3. Description of the sample: for instance the name of the facility;
4. *Location code*: is a sample point codification created in LIMS, which define the sample. It is composed by the following underscore-separated codification:
    a. Request identification number (RIN), which identifies the project;
    b. Sample type: codification used to describe the matrix of the sample. In this case, "SLF" stand for "freeze-dried sludge";
    c. Collection ID: it identifies in an intuitive way the sampling point;
    d. Moment ID: could be the time at which the sample was collected or, in case of more than one sample collected at the same location, a progressive number identifying each sample container;
    e. Depth: is the sampling depth. Identify, in the sample cores, the soil layer or the point in the water column. When a depth is not identified, for example in the case of bulk samples, the code 00 is used.

5.2. Entry results

Once results are ready, they must be validated, including both evaluation and formal approval. After validation they are archived in LIMS. This consent to track results of tests conducted in laboratories, which could be used for final reporting activities.

For each type of analytical methodology applied to the samples (i.e.: sample pre-treatment procedures), an analysis code is created adapting to the following format:

!_I_FDS_1_FD

The analysis code is composed by the following underscore-separated codification (from left to right):

a. Analysis type: express the type of the analysis ("$" = multi-component; "!" = text; " " = number) ;
b. Section ID: is the laboratory section where the sample is analysed. In the example "I" stand for 'inorganic';
c. Method ID: identify the method used for the sample analysis (e.g. Freeze-drying for sludge samples);
d. Variant: identify variations of a method (e.g. different parameter conditions for the same 'Method ID');
e. Instrument ID: it is the code used to identify the equipment (e.g. FD = freeze-drying system)

Chapter 6

Application 1- Gromo mine site

In this chapter PMF was applied to a local scale data set, considering an area of about 40.000 m^2. In this way, it was possible to combine PMF result with a GIS-based approach, for a better factors resolution.

The data set is characterized by the geochemical characteristics of the abandoned *Coren del Cucì* mine dump (Upper Val Seriana, Italy), which lead to waste rock accumulation due to ancient mining. Statistical methods are increasingly used for geochemical characterization of contaminated sites, particularly in order to understand which are the anomalies of natural and man-made and timely delivery to extend in two or three dimensions.

Abandoned mines are one of the most important environmental problems connected to mining activities (US EPA, 2000). In the European Union (EU), mining waste is known to be amongst the largest waste streams and it ranks first in the relative contribution of wastes in many Central and Eastern European Countries (Puura *et al*., 2002).

Abandoned mine sites consist of waste rocks that tend to accumulate in open pits, tailing and waste disposal areas. Their impact ranges from land degradation to abandoned waste disposal areas, which could be characterized by a residual mineralization and high metals content. In addiction, when minerals in abandoned mine sites are exposed to the weathering effects of air and water, acid mine drainage (AMD) may occur and result in release of metals into the surrounding environment (US EPA, 2000), posing a potential risk for water and soil systems.

Characterization of waste disposal areas is of great interest to assess their environmental impact (Puura *et al*., 2002). Identification of potential pollution sources or processes may be carried out by means of multivariate statistical approaches.

Multivariate statistical techniques are usually applied to geochemical data sets from waste disposal areas, to determine the number and composition of contamination sources, geochemical processes as well as hidden data structures (Kaplunovsky, 2005, Mostert *et al*., 2010). Moreover, the combination of multivariate statistical techniques with a geostatistical approach, such as variogram and kriging analysis, contributes to identify the impact point of resolved sources/processes (Schaefer *et al*., 2010).

PCA and PMF were used to investigate how different approaches deal with the preset type of data, while CA was used to extract two more homogeneous data subsets for PCA analysis. In

particular, a comparison between PCA and PMF results was carried out to highlight positive and negative aspects of their application. In addiction, ordinary kriging interpolation was applied to PMF resolved factor scores (**G** matrix elements) to visualize the potential environmental impact of the waste dump site.

6.1. Study area

The abandoned *Coren del Cucì* mine dump is located near the Gromo village (Upper Val Seriana, Italy). For details on geology, petrography and mineralogy of metal deposits of the *Coren del Cucì* area, readers are referred to Servida *et al*. (2010). In **Fig. 12** an aerial photo of the study area is reported. In the past, the mine was used for the exploitation of heavy metals, such as Fe, Cu, Pb, Zn and Ag (Jervis, 1881), confirmed also by the presence of numerous adits situated in the area. Nowadays, the mine area is comprised predominantly of waste rocks disposed over an area of about 40.000 m^2 (Servida *et al*., 2006). The waste disposal area is surrounded by vegetation (forests and grass). The grass field is situated mainly to the east of the waste disposal area (see **Fig. 13-a** for a view of sampling locations).

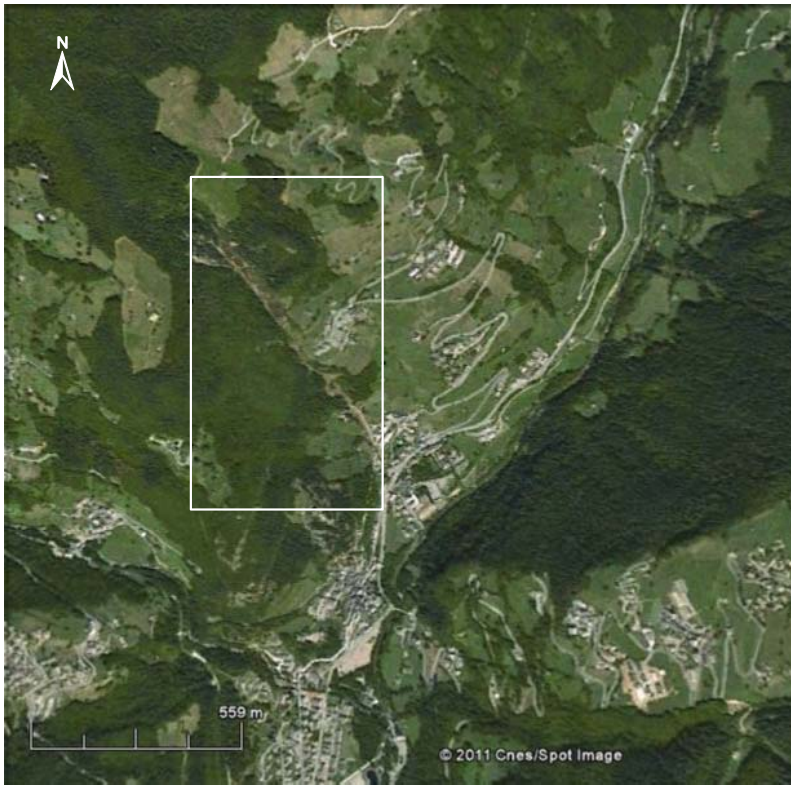


Fig. 12: Aerial photo of the Gromo mining site (from Google Earth). The white box indicates the study area, corresponding to the *Coren del Cucì* mine site

6.2. Data set description

The study data set consist of concentrations of some major elements (Ca, Fe, Mg), heavy metals (Ag, Cd, Co, Cu, Ni, Pb, Zn) and As, and values of pH in 56 samples of which only those present outside the dump are strictly classified as soil samples. The collection of samples, from both inside and outside the dump, was performed using the FOREGS sampling method (Salminen *et al*., 1998). The pH was determined using a pH-meter after suspending the soil in distilled water (ratio soil/water 1/2.5). For the analysis of major elements and heavy metals, samples were first grinded (<60 μm) and then digested with 6 ml 30% HCl Merck Suprapur and 2ml 65% HNO3 Merck Suprapur in a closed microwave oven (Milestone 1200 Mega), using the *aqua regia* method (ISO, 1995). Major elements and heavy metals concentrations were determined by ICP-AES (Jobin Yvon JY24) directly in solution. Concentrations of As were measured using the hydride method. Calibration for this element was done with the standard addition method. The chemical concentrations were measured in triplicate and the resulting percentage coefficients of relative standard deviation were below 10%.

Four main classes of samples were identified based on their locations in the examined area: dump and dump/forest, for samples collected inside the dump; and forest and grass, for samples collected outside the dump (**Fig. 13-a**)

Below-detection-limit data were identified by the notation '< DL' (detection limit) and no measured values, i.e. uncensored data, were reported. Although the use of uncensored data is preferred (Farnham *et al*., 2002), in this situation individual variables measured BDL were replaced by 1/2 the detection limit. Missing values were substituted with the mean value for each parameter.

For all the mentioned techniques, a modification of the pH parameter was applied before the statistical analysis. As expressed in Reinikainen *et al.* (2001), the expression 7.5-pH was used instead of the pH parameter, because it has the property that it increases when the acidifying emission increases.

Prior to PCA and CA analysis, outliers were detected using the Mahalanobis distance and were removed from the analysis. Moreover, variables with a high proportion (>5%) of below-detection-limit values and/or missing values were omitted from the analysis as they could strongly affect the results (Templ *et al*., 2008).

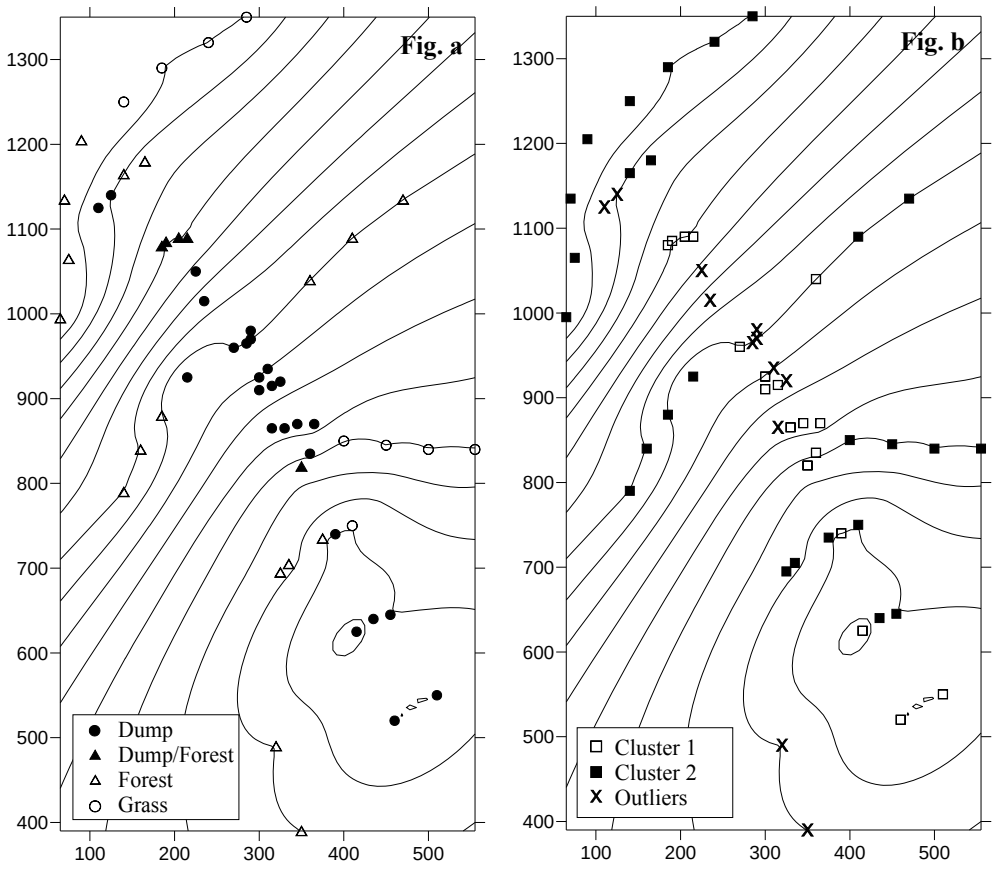


Fig. 13: a) Sample classification of the study area; b) representation of cluster analysis results and identification of detected outliers. The study area corresponds to the white box shown in Fig. **12**.

The influence of different normalization and standardization pre-treatment procedures on PCA outputs was examined. Two types of normalization procedure, logarithmic and Box-Cox transformations, were tested to take into account deviation from a normal distribution. Autoscaling (also called z-transformation) and Pareto scaling, similar to the former but using the square root of the standard deviation as scaling factor, were evaluated also.

6.3. Descriptive statistic

In the analyzed data set, BDL values of Ni and Ag comprise 2% and 14%, respectively, of all samples. Only the variable pH contains missing values, comprising 4% of all samples.

Descriptive statistics of measured elements and pH are given in **Tab. 4**. Boxplots of element data in logarithmic scale are shown in

Fig. 14.

The presence of different populations in the same data set (mine dump material, soils in forest and grass) is likely the reason for the high coefficients of variation (CV) of every variable (Errore. L'origine riferimento non è stata trovata.**)**. Moreover, the distributions of the majority of the elements, except Mg and Cd, are strongly positively skewed, with skewness coefficients > 1.

Tab. 4: Descriptive statistics of elements concentration (mg/kg) and pH parameter.

Element	***Ca***	***Fe***	***Mg***	***Zn***	***Cu***	***Pb***	***Co***	***Ni***	***Ag***	***Cd***	***As***	***pH***
Min	254	15935	786	43	3.4	37	6.1	4.3	0.1	0.4	9.5	4.1
Max	30371	84082	4438	19889	2861	7446	424	255	72	18	2093	6.3
Mean	5782	37173	2340	649	555	390	106	53	15	7.2	563	5.1
Median	1523	34441	2151	183	256	220	47	26	1.9	6.2	346	5.1
SD (±)	9924	12758	842	2638	675	1019	111	58	20	3.8	520	0.6
Skewness	1.82	1.26	0.66	7.11	1.46	6.30	1.17	1.68	1.16	0.35	1.33	-0.25
CV (%)	172	34	36	407	122	262	105	108	132	53	92	11

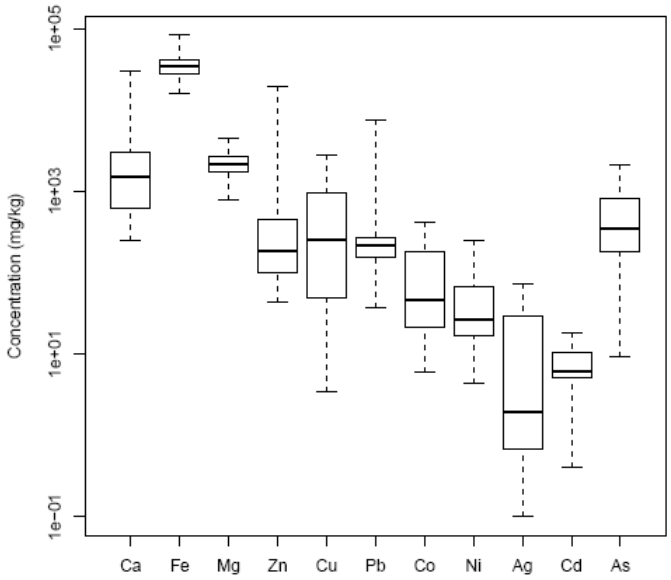


Fig. 14: Boxplot showing the variation of the measured elements concentration (mg/kg): median, 1st and 3rd quartiles and whiskers (lowest and highest values)

6.4. Cluster analysis

Cluster analysis was used as a prior step to cluster observations in order to extract more stable data subsets to be used as input for principal component analysis. In this way, grouping locations that show a similar behaviour, more suitable sub-groups of samples for PCA analysis were obtained. Logarithmic transformation and autoscaling were applied to the dataset. Cluster analysis was performed with R software (R Development Core Team, 2005) using Ward hierarchical agglomerative method with Euclidean distance.

According to the data pre-treatment procedures described above, Ag data were omitted from the analysis because >5% of the values were BDL.

The fist two main clusters, resulting from the analysis, were selected as two independent data sets to further separate examination by PCA technique. In **Fig. 13-b**, a graphical representation of resolved clusters is given, showing sample-cluster association; samples classified as outliers are also shown. Sampling sites belonging to cluster 1 are those located in the waste disposal area, while sampling sites belonging to cluster 2 are associated with the forest and grass areas surrounding the dump. The elements Ag, As, Co, Cu and Ni show higher average concentrations in cluster 1 than in cluster 2, confirming their association to the dump area. No significant variations were observed for the other elements and the parameter pH. It is also pointed out that most outlier values of some elements pertain to the dump zone.

6.5. Principal Component Analysis

PCA was conducted on the resolved clusters separately. Indeed, it is important to underline that PCA gives optimum results when applied to homogeneous sub-populations separately (Reimann *et al*., 2002); its application to heterogeneous data may results in a distortion of principal components. R software (R Development Core Team, 2005) was used to perform PCA based on the singular value decomposition (SVD) algorithm. Principal components with eigenvalue greater than 1 were selected (Kaiser criterion).

Since two distinct populations were evidenced by cluster analysis, it was chosen to apply PCA on the two populations separately, inside and outside the dump, made by 27 and 25 samples, respectively. The chosen pre-treatment procedures for both the two analyzed sub-sets were logarithmic transformation with Pareto scaling, according with a better possible explanation of PCs extracted.

6.5.1. Area inside the dump

Three samples were eliminated as statistical outliers. The pH data were omitted from the analysis because more than 5% of the values were missing.

Three PCs were extracted, explaining about 80% of the cumulative variance. Scatter plots of PC1 vs. PC2 and PC1 vs. PC3 are shown in **Fig. 15**.

The first component, explaining 46% of the total variation, is characterized by positive loadings for Ag, Cu, Co, Ni, and As. According with the localization of the analyzed sub-population inside the dump, PC1 could be identified with the mineralization matching the ores located in the mining area. More in detail, chalcopyrite, native silver, arsenopyrite and Co-Ni sulfarsenides were found in the considered area (Servida *et al*., 2010). The PC2 is determined by positive loadings for Ca and Zn and, to a lesser extent, for Ni. This component covers 20% of the total variance. Calcium and zinc could be attributed to a background component. In particular Ca may be connected with the non-mineralized substrate, for example micaschists and carbonates of the outcropping rocks, and Zn with the sulphide bearing minerals not bound to the main mineralization which conditions the presence of elements in the dump materials. This is supported by the fact that sphalerite, the main zinc sulphide, was not detected as ore mineral assemblages in the mine area (Servida *et al*., 2010). Finally, PC3 was strongly dominated by cadmium. This component, which accounted for 12% of the total variance, could be associated with a high natural background concentration of Cd. Indeed, Cd showed the lowest coefficient of variation (**Tab. 4**), with a constant concentration distribution over the whole area.

Results provided by the other tested transformations, used to investigate the effects of data pre-treatment methods, are here summarized. Autoscaling, with both Box-Cox and logarithmic transformations produced slightly different PCs. The Mg was explained by PC1 and, in general, loadings were lower in all the PCs extracted. Using Box-Cox transformation with Pareto scaling, more than 80% of variation was explained by the PC1, which was determined by high positive loadings for Mg only.

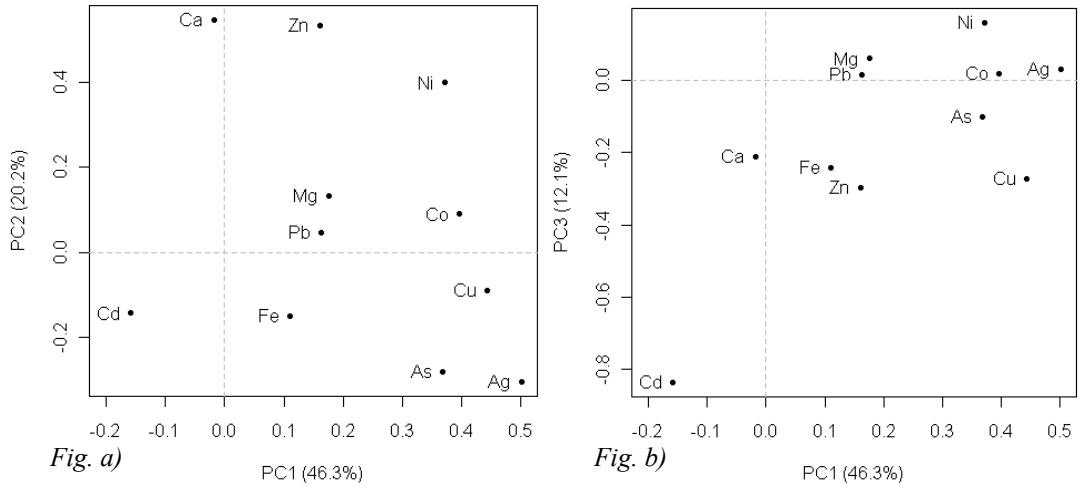


Fig. 15: plots of PCs extracted from PCA applied to the sub-population of samples located inside the dump; *a*) PC1 vs. PC2; *b*) PC1 vs. PC3. The amount of the explained variance is indicated in brackets.

6.5.2. Area outside the dump

One sample was eliminated as outlier. The Ag data were omitted from the analysis because>5% of the values were BDL.

Two PCs were extracted, explaining about 70% of the total variation. Scatter plots of PC1 vs. PC2 is shown in **Fig. 16**.

The first principal component, accounting for 48% of the total variance, was positively correlated with calcium and, to a lower extent, with Zn and Pb. PC1 could be attributed to the background component dealing both with the non-mineralized substrate, together with Zn and Pb sulphides localized outside the dump (Servida *et al*., 2010).

The PC2, explaining 20% of the total variance, is characterized by high-positive loadings in Cu and moderate-positive loadings in Co and As. These elements can be associated with the residual mineralization which extends outside the dump site (Servida *et al*., 2010). The intermediate position of the remaining variables may indicate a joint contribution from both a natural source and mineralization.

Results provided by the other tested transformations, used to investigate the effects of data pre-treatment methods, are here summarized. Autoscaling, with both Box-Cox and logarithmic transformations, resulted in a PC1 characterized by negative loadings for Cd, Fe, Mg, Ni, Cu and pH. The PC2 reflected the above mentioned results, but showing lower loading contributions. Using Box-Cox transformation with Pareto scaling, more than 80% of variation was explained by PC1, which was characterized by high negative loadings for Mg only.

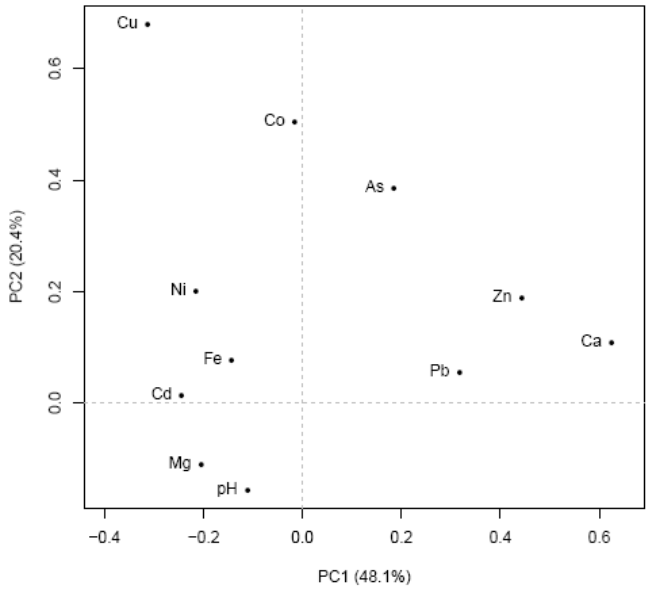


Fig. 16: Plot of PCs extracted from PCA applied to the sub-set of samples located outside the dump; the amount of the explained variance is indicated in brackets

6.6. Positive Matrix Factorization

The program PMF2 (Paatero, 1997), version 4.2, was used to solve the two-way PMF model. PMF analysis was performed using the *robust mode* with an outliers distance equal to 4. From 2 to 8 factor solutions were investigated with the *FPEAK* parameter ranging between -1 and +1 (Reff *et al*., 2007) with a 0.1 incremental step.

Error estimates used to weight data were computed by means of the EM=-14 error model structure, implemented into the algorithm. This option, recommended for general-purpose environmental work, computes the standard deviation matrix (s_{ij} matrix elements) according to the following equation (Paatero, 2007b):

$$s_{ij} = t_j + v_j \cdot \max\left(\left|x_{ij}\right|, \left|y_{ij}\right|\right)$$

x_{ij} are the elements of the input data matrix and y_{ij} are the fitted values; t_j and v_j are parameter coefficients computed as following. Typically, in environmental work, the t_j values equal the detection limit of each variable.

Since in the study data set the detection limits were known only for Ni and Ag, the min(x_j)/4 values computed for the remaining variables were used for t_j estimation in the standard deviation matrix. The v_j coefficients were chosen by trial and error using the Q value as optimization parameter (Polissar *et al*., 2001). Values for t_j and v_j parameters are given in **Tab. 5**.

Tab. 5: t_j and v_j values used in the EM=-14 error model equation.

	Ag	*As*	*Ca*	*Cd*	*Co*	*Cu*	*Fe*	*Mg*	*Ni*	*Pb*	*Zn*	*pH*
t_j	0.1	2.36	63.5	0.1	1.53	0.86	3984	196	4.3	9.20	10.8	0.298
v_j	0.16	0.08	0.15	0.04	0.15	0.1	0.03	0.04	0.09	0.18	0.25	0.8

In order to obtain larger error estimates for BDL and missing values, the v_j coefficient was multiplied by 2 and 4, respectively.

The selection of the optimum solution was based on the analysis of Q values obtained in different runs, varying the number of factors and the rotational degree. In addition, for improved results, the output parameters *RotMat*, *IM*, *IS* (Lee *et al*., 1999) were inspected.

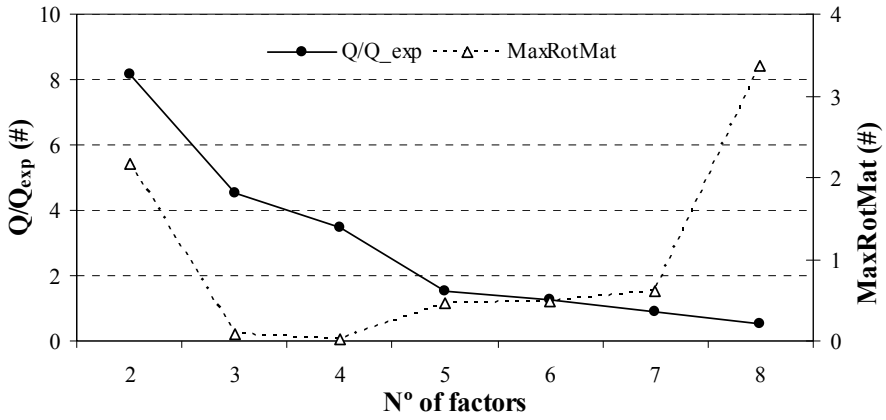


Fig. 17: Q vs. Q expected (left) and RotMat (right) parameters for each number of factors examined

The Q/Q_{exp} ratio determined for each analysed number of factors is shown in **Fig. 17**. Since for the 7-factor and 8-factor solutions model the ratio assumes a <1 value, solution with 7 and 8 factors extracted were rejected. In the same figure, also MarxRotMat values are given; it can be observed that they assume lower values for a number of factors between 3 and 7.

IM and IS parameters are illustrated in **Fig. 18**. Their values rapidly decrease when 3 factors were resolved, with a further decrease for the 5-factor model. The range of optimal solution could thus be restricted from 3 to 6 factors. However, looking at the not explained variation values, given in **Tab. 6**, Ca is not explained by the 3 and 4-factor solution (high NEVF). In addition, Zn shows a decrease in its NEVF in the 5-factor model, explaining a component defined by Zn and Ca variations. For these reasons, solutions with 3 and 4 factors resolved were rejected.

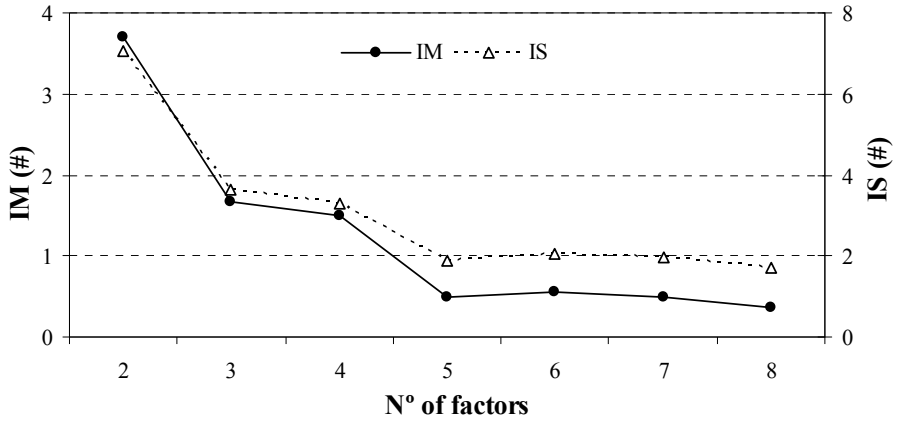


Fig. 18: IM and IS parameters values for each examined number of factors.

Tab. 6: NEVF for different number of factors in the central solution (FPEAK=0).

Factors	*Ag*	*As*	*Ca*	*Cd*	*Co*	*Cu*	*Fe*	*Mg*	*Ni*	*Pb*	*Zn*	*pH*
3	34%	0%	55%	2%	23%	13%	18%	27%	30%	36%	46%	28%
4	29%	5%	46%	5%	21%	9%	9%	15%	15%	27%	39%	19%
5	26%	3%	3%	3%	22%	7%	10%	14%	21%	27%	26%	18%
6	3%	0%	5%	1%	18%	6%	13%	17%	10%	27%	26%	18%

Examining the rotations influence for the 5 and 6-factor models, with FPEAK parameter ranging from -1 to +1, the obtained results did not differ significantly from the central solution, in terms of explained variation. The difference between the 5 and 6-factor solutions is only given by the explanation of silver in a unique factor in the 6-factor solution,. Since no clear interpretation was found for silver variation, the 5-factors solution was chosen as the more representative. Considering the Q_{rot}/Q_{cent} ratio (**Fig. 19**), rotations with FPEAK greater than 0.5 were discharged, because they show a >5% difference between rotated and central Q. These rotations also show higher IM and IS values (**Fig. 20**).

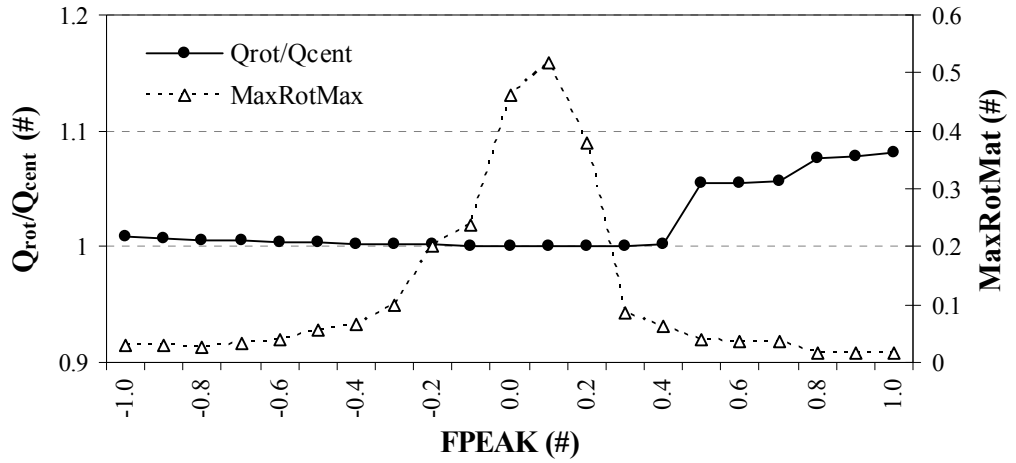


Fig. 19: Q for rotations vs. Q for central solution (left) and RotMat (right) parameters varying the FPEAK value.

Since no significant changes in factors resolution were observed for the remaining rotations, the central solution was chosen. Explained variations for the 5-factor solution, expressed in percentage terms, are shown in **Fig. 21**. Spatial distribution maps of factors, illustrated in **Fig. 22**, were obtained applying ordinary Kriging interpolation on the factor score matrix **G**. Factors maps were used in helping to understand the PMF factors interpretation.

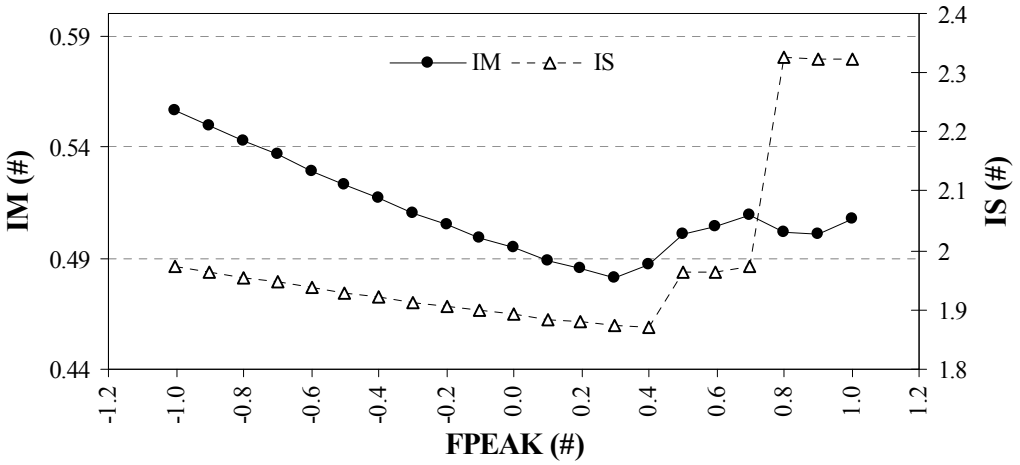


Fig. 20: IM and IS parameters for different FPEAK values tested.

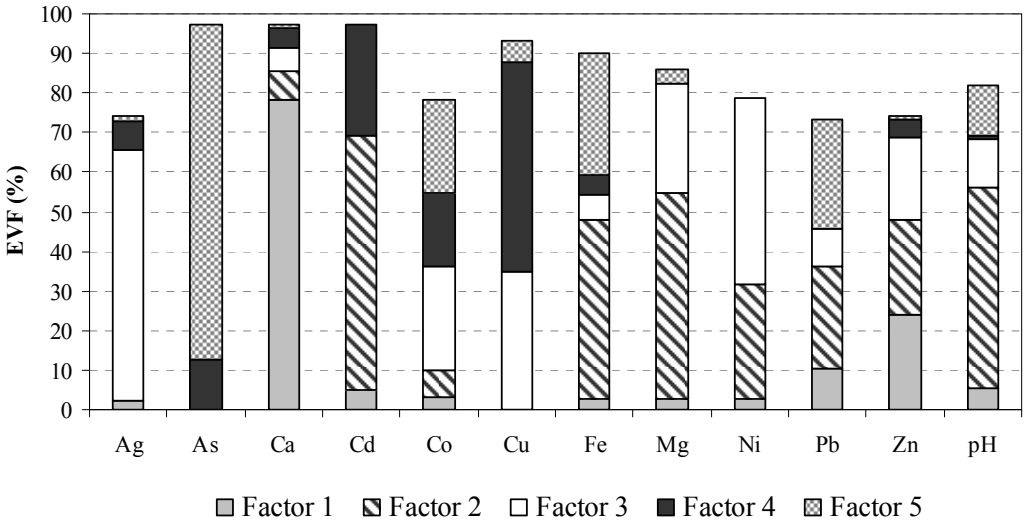


Fig. 21: Explained Variations of F matrix for the 5-factors solution with FPEAK=0.0

Factor 1

Factor 1, which is mainly characterized by Ca variation (78%) could be interpreted as the non-mineralized substrate. Indeed, Ca was found to be the major component of the main outcropping rocks both of silicate Crystalline Basement and of carbonate Mesozoic cover in the study area (Servida *et al*. 2010). Zn, which is present in the regional geology, may also be included in this source, because its variation is higher compared to the other variables EVFs. However, Zn spreads its variation also in factors 2 and 3, being a typical element of the sulphur mineralization in the mine site (Servida et al., 2010). Examining the factor map (**Fig. 22-a**), high **G** values are mainly located in the grass and forest areas. An exception is found to the south of the waste disposal area, which is however characterized by grass-like vegetation. Such an identification allow to associate the non-mineralized substrate with the grass area and forest that grow on soils, which here make the sampled materials and, as well known, calcium is a major component of soils (Mitchell, 1964).

Factor 2

The variability of Fe (45%), Mg (52%), Cd (64%) and pH (51%) is explained by factor 2. It is pointed out that the Acid Mine Drainage (AMD) process, resulting from mining activity, was not observed in the study area (unpublished results). The variables explained by factor 2 show the lowest coefficients of variation over the whole mine site, indicating a lower variability compared to the other measured variables. Factor map, illustrated in **Fig. 22-b**, exhibits a homogeneous distribution across the mine site, except in the waste disposal site characterized by a residual mineralization (see factor 3). This suggests that factor 2 may be associated with a component

controlled by parent rocks. This is in accordance with the geological and mineralogical characterization of the considered area given by Servida *et al*. (2010). Indeed, Fe attends both in rocks and in mineralizations, particularly as siderite ($FeCO_3$) that is the main mineral disseminated on the entire area; Mg pertains to substrate materials and is a main component of soils; and Cd may be found as a minor component in the sphalerite structure that is localised prevailing in the area outside the dump. Moreover, at the pH values here detected, cadmium exhibits a higher mobility respect to the mobility characteristic of element forming ore phases (Chuan *et al*., 1996; Kabata-Pendias and Pendias, 2001).

Lead spreads its contribution in both factor 2 and factor 5. Typically, it occurs in the galena mineralization, which is found to be mainly associated Fe-containing minerals (chalcopyrite, sphalerite and sulphuarsenides; Servida *et al*., 2010).

Factor 3

Factor 3 is characterized by Ag (63%) and Ni (47%) variability and less strongly by Cu (35%). According to Servida et al. (2010) the ore mineralization in the mine site is represented by a variety of sulphides and sulphosalts containing among others also Ag, Ni and Cu. The factor 3 distribution is localized along the dump zone (**Fig. 22-c**), suggesting a connection with the ore mineralization characteristics of the area inside the dump. The mineralization also includes the Co, here explained in a percentage <30%. However, Co spreads its variation also in factors 4 and 5, indicating a common source of mineralization.

Factor 4

Factor 4 is determined by Cu variation (53%). Although it seems that this factor could be combined in the ore mineralization identified by factor 3, the PMF 4-factor solution did not produce a satisfactory result. With four resolved factors in fact, more than 45% of calcium variability was not explained by the model. The spatial distribution map of factor 4 (**Fig. 22-d**) shows a high impact zone in the northern part of the dump area and a moderate impact in the central part of the dump.

This distribution could be compatible with the presence of two of major Cu-bearing minerals, chalcopyrite and tetrahedrite, both occurring in nearness of the adits and along the dump (Servida *et al*., 2010). Factor 4 is in close relation with factor 3, being copper also included in the sulphide mineralization explained by factor 3.

Factor 5

Factor 5 is characterized by As variation (95%) and, to a lower extent, by Fe (30%). High **G** scores are distributed in the central part of the waste disposal area and, with a minor extent, close to the north and south edges of the dump site (**Fig. 22-e**). This suggests a localized anomaly of arsenopyrite, characteristic ore phase of the Coren del Cucì dump (Servida *et al*., 2010), but not the exclusive for the presence of As that is a component also of the other ore phases found as tetrahedrite and sulphoarsenides. Moreover, a correlation coefficient 0.63 between As and Fe for samples collected inside the dump, indicates the relationship between these elements exclusively in ore minerals on the dump. No correlation is found outside the dump, confirming the characterization of iron given in factor 2.

The factor 5 spatial distribution map, displays an opposite trend respect to factor 4, confirming the occurrence of two distinct geochemical anomaly zones.

6.7. Conclusions

Results provided by PCA for the sub-population located inside the waste disposal area describe a source of mineralization, together with a possible geo-mineralogical component characterized by a high natural background value for cadmium. Outside the dump, a residual mineralization component was explained by positive loadings for Cu, Co and Ni. Moreover, for both the examined sub-populations, a common source connected with the non-mineralised substrate and main Zn sulphides was determined. No particular and interesting information or hidden data structures were extracted from PCA analysis.

The application of the PMF approach lead to more interesting results, supported also by the fact that a GIS-based technique was successfully combined with the positive PMF scores produced. Five factors were resolved. Two well separated background components were distinguished outside the dump area, matching with the non-mineralized substrate (similarly to PCA results) and with parent rocks characterization. A main component, explaining the ore mineralization inside the waste disposal area was identified by Ag, Ni, and Cu variations. However, the more interesting factors were two geochemical anomaly zones characterized by copper and arsenic mineralization, respectively.

In conclusion, PMF was found to be a useful tool for the characterization of abandoned mine sites, being able to identify mineralized components, i.e. geochemical anomalies. Moreover, the

combination with a GIS-based approach was successfully used to identify the impact point of the resolved sources.

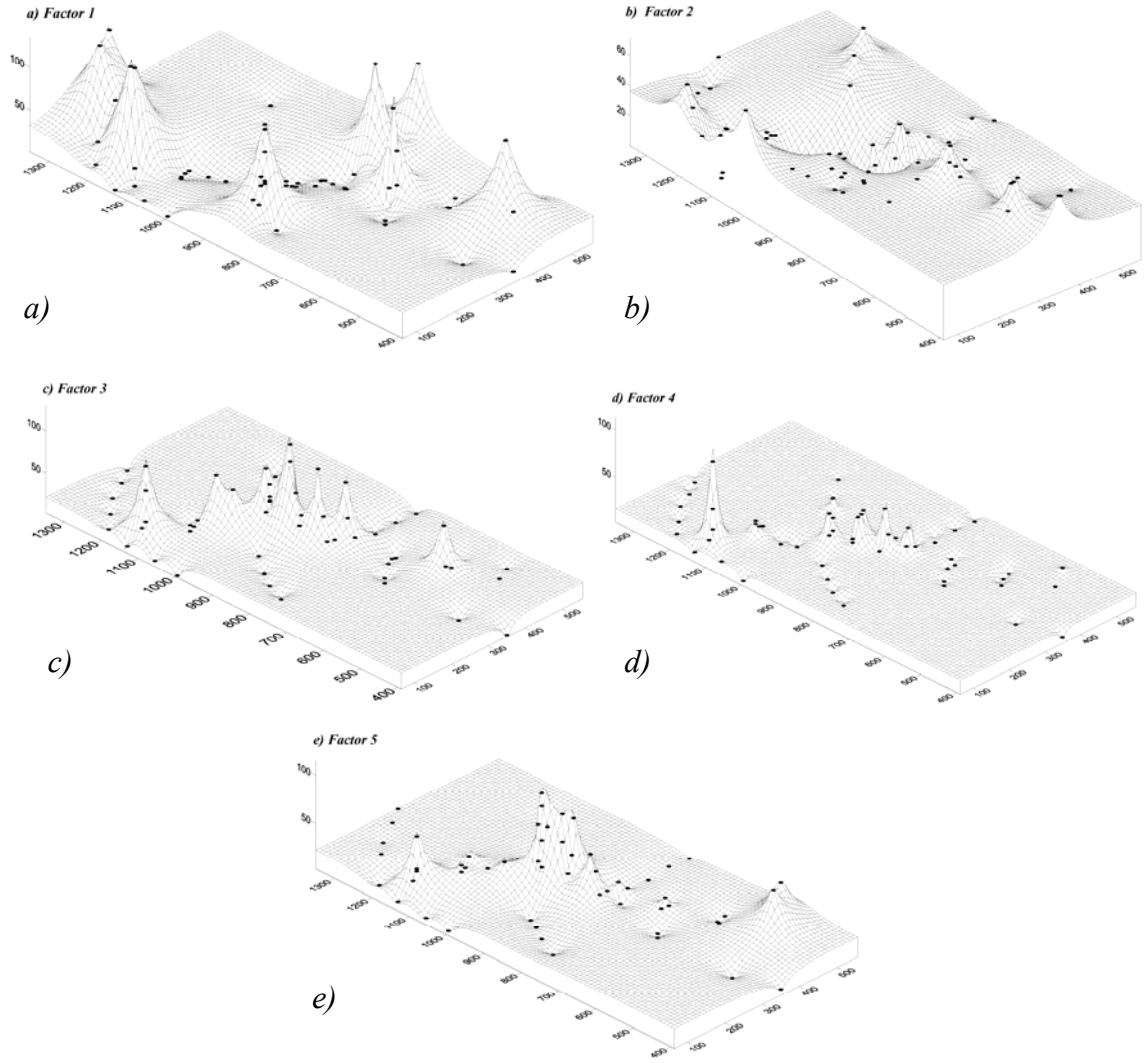


Fig. 22: spatial distribution maps of PMF resolved factors computed using ordinary Kriging interpolation. Scale is in meters distance.

Chapter 7

Application 2 - Alpine lakes

In this chapter, positive matrix factorization was applied in the context of a pan-regional study characterized by sub-populations of samples affected by different geological features. In particular, the study focused on the characterization of alpine lakes located in the northern part of Italy. The data set is represented by sub-populations of sediment samples collected at eleven different lakes. The sediments samples were collected within the frame of the project "*An ecological assessment system for sub-alpine lakes using macroinvertebrates – The development of a parsimonious tool for assessing ecological health of European lakes*" funded by the Technology Transfer and Scientific Cooperation Unit. The purpose of the project was to examine the importance of environmental factors, among which sediment chemical characteristics, that can affect macroinvertebrate communities. In particular, the evaluation of sediment chemical characteristics was used to evaluate the relative role of sediments in explaining macroinvertebrate abundance.

The PMF approach applied on lakes sediments samples aimed at the determination of main factors which explain sediments composition, including the possibility to discover contamination sources. Factors identification, performed by PMF, was compared with results obtained by the two most common multivariate techniques: principal component analysis (PCA) and cluster analysis (CA).

7.1. Data set description

The study data set contains chemical composition data obtained in sediments samples from 11 alpine lakes located in Northern Italy.

Sediment samples (100 g) were taken from the sub-littoral zone of each lake stations using an Ekman grab. They were dried at 40 ºC and then sieved through a 2-mm mesh and ball-milled. For each lake, 17 to 20 samples had been collected, with a total of 196 samples (**Fig. 23**). A total of 21 elements were measured by a wavelength-dispersive X-ray fluorescence (SRS-3400, Bruker-AXS®): Al, As, Ca, Cd, Cl, Co, Cr, Cu, K, Fe, Mg, Mn, Na, Ni, P, Pb, S, Si, Ti, V and Zn. For further information on the analytical methodology as well as regarding the analytical quality control measures taken, refer to Free *et al*. (2009).

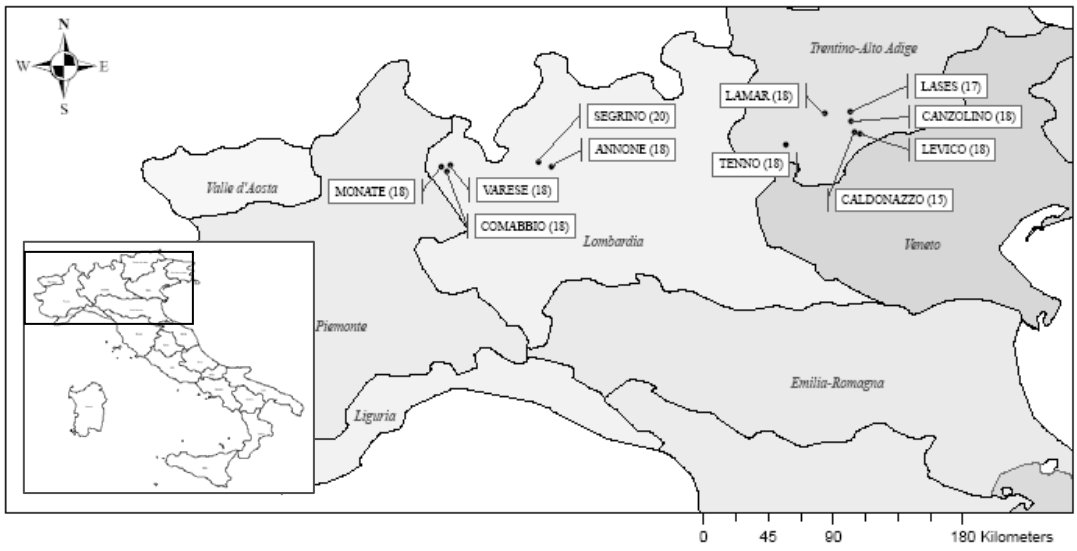


Fig. 23: location map of the examined lakes. The number in brackets is the number of samples collected.

7.2. Descriptive statistic

In the examined data set, below detection limit data were identified by the notation '< DL' (detection limit) and no measured values (i.e. uncensored data) were reported in such a situation. Hence, concentrations at or below the respective limit of detection were censored by replacement with ½ the DL concentrations. No missing values were found in the data set. Cd was omitted from the analysis because all the concentrations were BDL.

Descriptive statistics (min, max, mean value, standard deviation and coefficient of variation) and percentage of BDL values are listed in **Tab. 7**; box-plots of element concentrations are shown in **Fig. 24**.

Positive skewness was found for the majority of the measured elements except for Si, S, Ca, Ti and V. Large coefficients of variation, in the range 52% - 200% were found for all the parameters. This could be attributed to the different geological features of the lakes, which are conditioned by the native mineralogy of the sediment.

Tab. 7: Summary statistics for the measured elements (wt % = weight percentage).

Element	*Min*	*Max*	*Mean*	*SD*	*CV %*	*% BDL*	***Skewness***
Na(wt%)	0.13	2.04	0.57	0.47	83	-	1.3
Mg(wt%)	0.29	7.20	1.33	1.28	96	-	2.8
Al(wt%)	0.20	15	3.86	2.78	72	-	1.0
Si(wt%)	1.87	35	17	8.88	52	-	0.1
P(wt%)	0.02	0.29	0.08	0.05	56	-	1.4
S(wt%)	<0.01	1.61	0.43	0.31	72	1	0.7
Cl(mg/Kg)	18	1322	117	119	102	-	7.4
K(wt%)	0.19	4.92	1.43	1.02	71	-	1.0
Ca(wt%)	0.28	30	12	9.52	77	-	0.2
Ti(wt%)	0.02	0.56	0.17	0.11	61	-	0.6
V(mg/Kg)	8	137	49	27	55	-	0.9
Cr(mg/Kg)	<2	259	50	49	98	19	1.6
Mn(mg/Kg)	43	1958	355	278	78	-	2.1
Fe(wt%)	0.42	6.68	2.17	1.37	63	-	1.0
Co(mg/Kg)	1	28	6.79	4.72	69	-	2.0
Ni(mg/Kg)	<2	180	12	23	192	59	3.9
Cu(mg/Kg)	<2	456	50	79	158	28	2.6
Zn(mg/Kg)	18	1162	186	174	93	-	2.4
As(mg/Kg)	<7	213	19	30	155	56	3.1
Cd(mg/Kg)	<9	-	-	-	-	100	-
Pb(mg/Kg)	27	3218	184	369	200	-	5.3

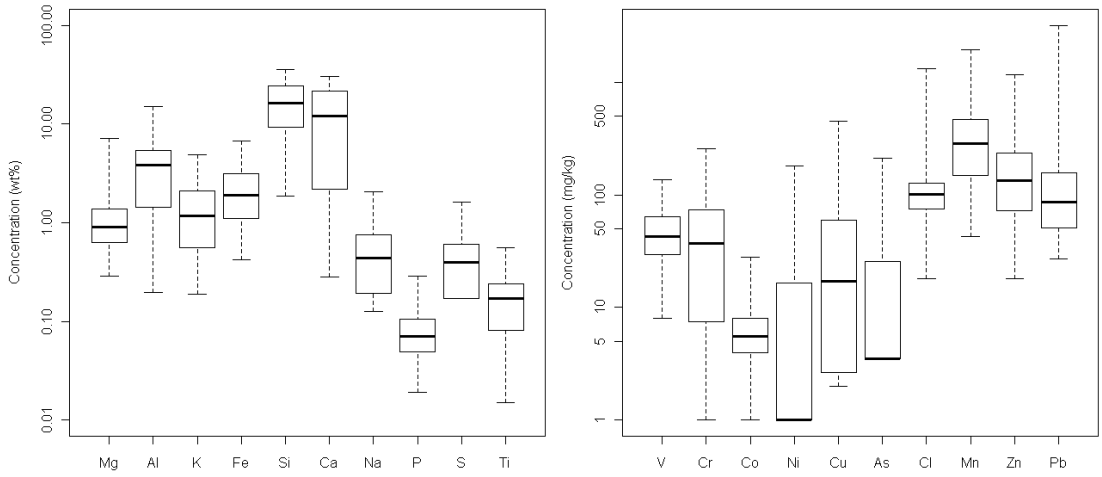


Fig. 24: Boxplots of concentrations of measured elements: median, 1st and 3rd quantiles, and whiskers (lower and highest values). The y-axis is plotted in logarithmic scale.

7.3. PMF analysis

PMF analysis was carried out using the *robust mode* with an outliers distance equal to 4. Solutions ranging from 2 to 10 factors were investigated with the *FPEAK* parameter ranging between -1 and +1 with a 0.1 incremental step.

Two different error estimates were tested to show possible variation in the resolved factors. Since measurement uncertainties were not available, two formulas found in literature were used.
The first type of tested errors structure, used by Xie and Berkowitz (2006), assigns higher errors to below-detection-limit data and was computed using the following equation:

$\sigma_{ij} = DL_{ij}/3 + d_j \cdot x_{ij}$ for representative data

$\sigma_{ij} = 5/6 \cdot DL_{ij}$ for below-detection-limit data

where x_{ij} is the *j*-element concentration at the *i*-location, and d_j are the element percentage parameters; d_j values, reported in **Tab. 8** were chosen by trial and error using *Q* value as optimization parameter.

Tab. 8: d_j percentage parameter values used in the Xie and Berkowitz equation.

Element	*Na*	*Mg*	*Al*	*Si*	*P*	*S*	*Cl*	*K*	*Ca*	*Ti*
d_j	0.1	0.1	0.07	0.1	0.1	0.1	0.1	0.07	0.05	0.05

Element	*V*	*Cr*	*Mn*	*Fe*	*Co*	*Ni*	*Cu*	*Zn*	*As*	*Pb*
d_j	0.07	0.1	0.15	0.07	0.07	0.2	0.15	0.1	0.15	0.1

The second error structure was derived from the work of Ogulei *et al*. (2006) and it was tested to account for the data variability:

$$\sigma_{ij} = k \cdot \left(x_{ij} + \bar{x}_j\right)$$

where $\bar{x}_j$ is the arithmetic mean of the *j*-element concentration and *k* is a multiplicative factor. The *k* factor was set equal to one tenth the relative standard deviation (RSD/10), to better reproduce the data dispersion. Moreover, this error structure gives large error estimates to small concentrations.
From different tests, initially computed with *FPEAK* set to 0 (central solution), no significant changes in the factor structure were observed by changing the error estimates; only little differences were found in the explained variation values of F. Finally, equation derived from Ogulei *et al*. (2006) was chosen to determine the optimal solution, in terms of the number of factors and rotations that better describe the problem under analysis.
Quality of fit was examined by means of *Q* values and scaled residuals obtained in different runs, varying the number of factors and the rotational degree. In addition, for improved results, *RotMat*, *IM*, *IS* and *G*-space plots results were inspected.
The first examined parameters were Q and RotMat (**Fig. 25**), and IM and IS (**Fig. 26**) in relation to the number of factors examined for the central solution (*FPEAK*=0).

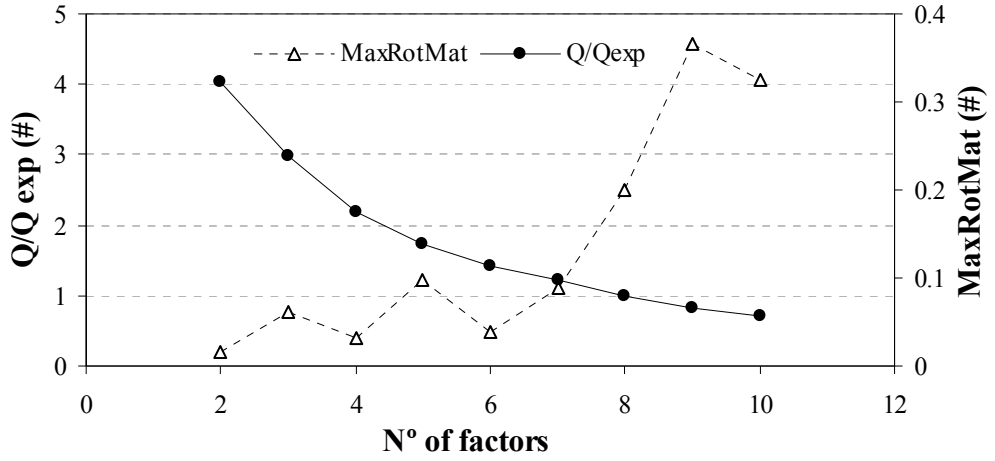


Fig. 25: Q vs. Q expected (left) and RotMat (right) parameters for each number of factors examined.

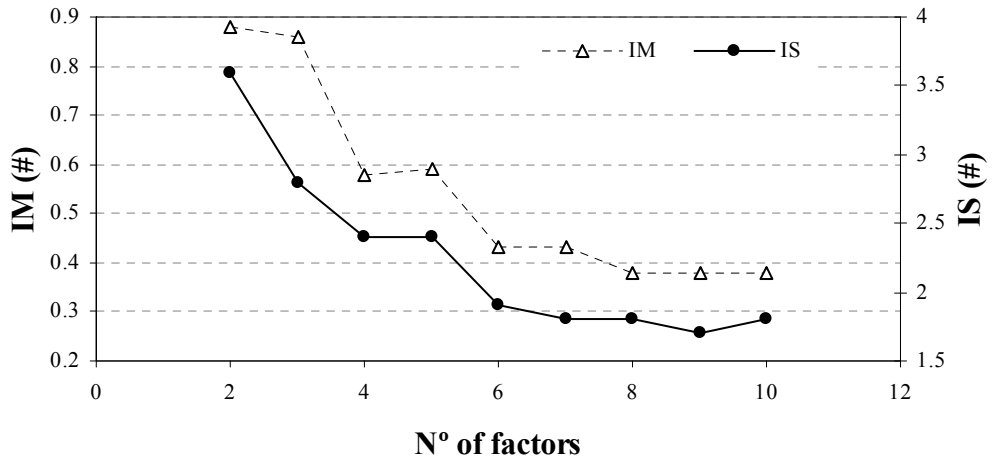


Fig. 26: IM and IS parameters values for each examined number of factors.

From **Fig. 25**, a gradual decrease of Q values can be observed, until it becomes equal to the expected Q value at eight factors resolved. Solution with more than 8 factors could be rejected because $Q/Q_{exp}<1$. Even the solution with 8 factors could be rejected since an increase of the MaxRotMat values occurs. Examining IM and IS parameters, a first decrease can be observed at the 4-factor solution, followed by a further decrease at six factors extracted. For these reasons, it was chosen to further examine only the solutions with 4 to 7 factors explained. NEVF were considered to compare the selected solutions (**Tab. 9**).

From **Tab. 9** it can be observed that NEVF significantly change for P, S, K and Cr passing to the 5-factor solution, while Mn and V show a decrease in the 6-factor solution. With 7-factors

identified, P and Mn reduce their unexplained variation, being uniquely explained by one additional factor.

Tab. 9: NEVF(%) for 4 to 7 PMF factors.

	Number of factors			
	n. 4	***n. 5***	***n. 6***	***n. 7***
Al	11	9	8	8
As	53	52	52	52
Ca	9	8	8	9
Cl	22	22	22	22
Co	14	15	15	15
Cr	**35**	**24**	23	21
Cu	53	51	48	49
Fe	9	9	8	8
K	**19**	**12**	12	10
Mg	26	26	25	23
Mn	28	**28**	**10**	6
Na	18	18	16	16
Ni	60	60	61	61
P	**18**	**13**	13	3
Pb	30	29	29	28
S	**19**	**12**	13	10
Si	13	11	11	11
Ti	14	12	11	11
V	14	**12**	**9**	9
Zn	23	23	23	23

The 6 and 7-factor solution attribute Mn to a unique factor; this could be mainly due to a high number of factors chosen, rather than to a new meaningful factor resolved. Solutions with 4 and 5 resolved factors differ for the explanation of Cr, P and S, which are grouped in a single factor in the 5-factor solution. However, no meaningful source was determined for their variation, which remains unaltered even exploring the rotational ambiguity. Therefore, the solution with 4 resolved factors was chosen as the most representative.

The source identification, in terms of explained variations (EVF), was also performed examining the rotational degree varying the *FPEAK* parameter. In this case, the Q value for rotations (Q_{rot}) was compared with the Q value obtained for the central solution (Q_{cent}).

In **Fig. 27** the Q value for the rotated solutions do not differ significantly (less than 1%) from the Q value obtained with FPEAK=0. However, the rotational ambiguity seems to be stronger for rotations closed to the central solution, where the MaxRotMax parameter shows higher values. Opposite to this behaviour, IM and IS (**Fig. 28**) show minimum values around the central rotation. Combining these results, it appears that the best fit is obtained for one of the following rotations: -0.5, -0.4, -0.3, 0.3 and 0.5.

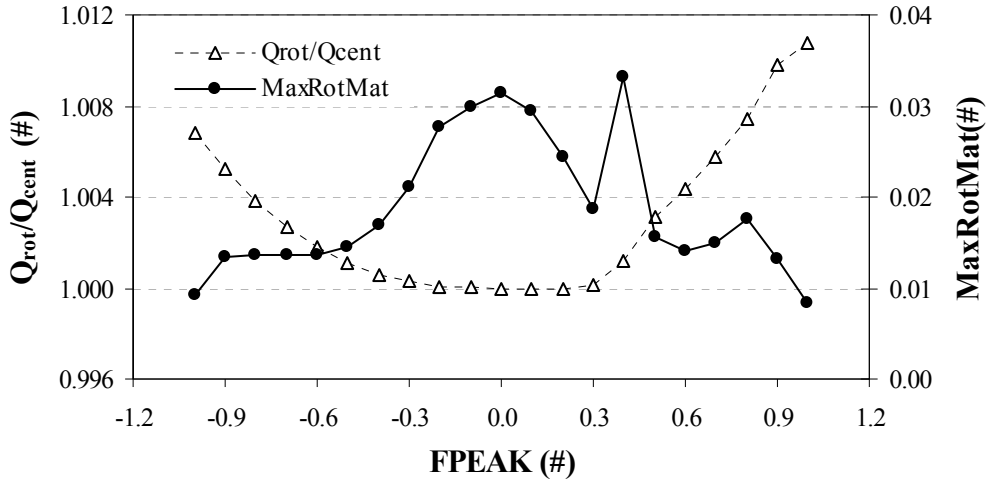


Fig. 27: Q_{rot} vs.Q_{cent} (left) and MaxRotMat (right) parameters for different *FPEAK* values.

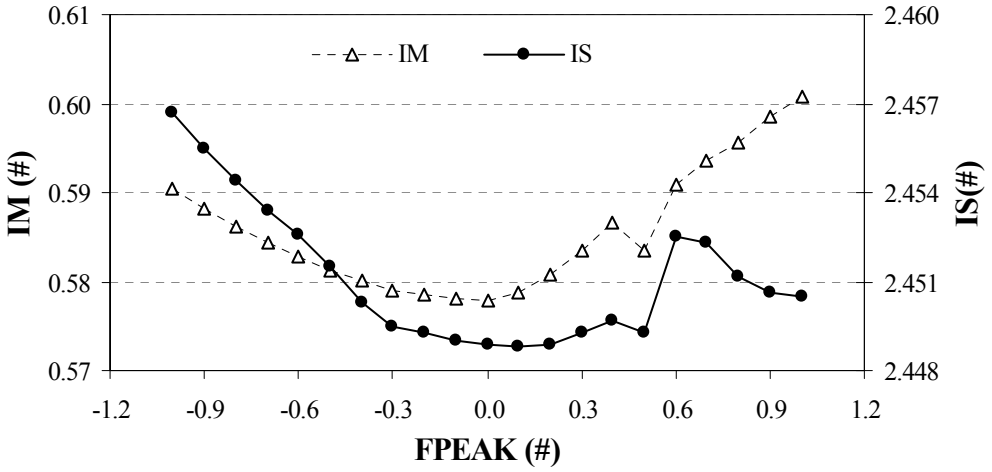


Fig. 28: IM and IS parameters for different FPEAK values.

To select the optimal rotation, G-plots were examined. However, plots show an analogous trend for all the examined rotations.

In **Fig. 29**, an example of G-plot is reported. It could be observed that, in the selected case, the resolved factors are independent each other.

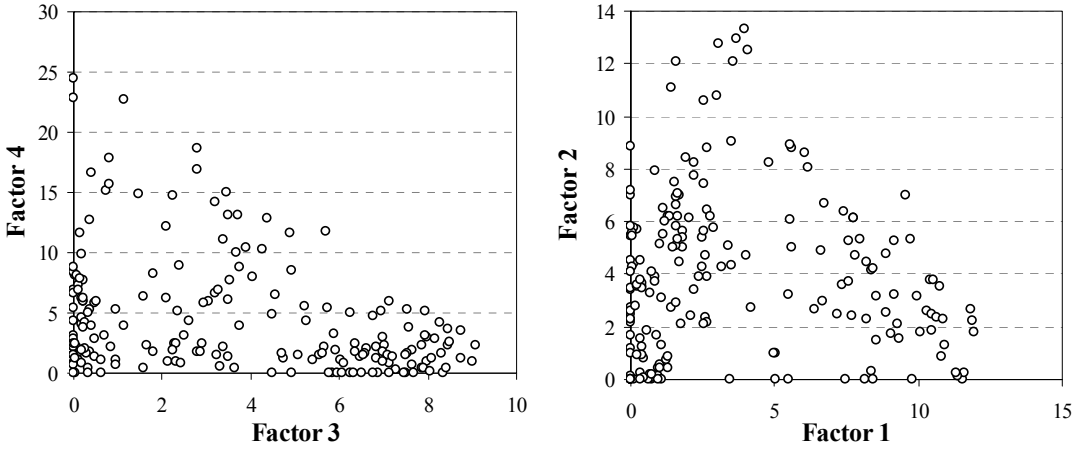


Fig. 29: Examples of G-plots for the 4-factor solution with FPEAK=-0.3.

Finally, the 4-factor solution with *FPEAK* parameter equal to -0.3 was chosen. Explained variations, used to identify the resolved factors, are shown in **Fig. 30**.

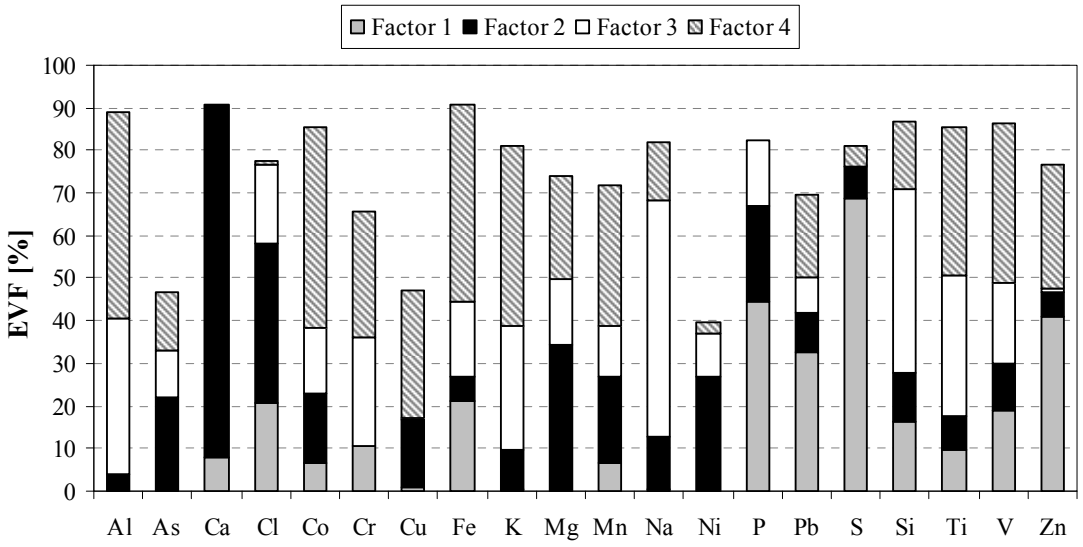


Fig. 30: Explained variations of F for the 4-factor solution with *FPEAK*=-0.3

Factor 1. This factor explains 70% of sulphur variation and, to a lower extent, Pb, Zn and P variation. Factor 1 was interpreted as a phosphate and sulphate/sulphide source. The presence of Zn and Pb, which have higher explained variations in this factor, could be associated both to sphalerite and galena, main zinc and lead sulphides, or to natural weathering processes of Zn-Pb-bearing minerals (Zaharescu *et al*., 2009).

Factor 2. The second factor accounts for most of the Ca variability (>80%) and could be related to a carbonate mineral source (for example calcite). This factor was also characterized by Mg

and Cl with about 30% of explained variability. The presence of Mg could be attributed to magnesium-carbonate ores (like dolomite), while no easily explanation could be given for the Cl element.

Factor 3. Factor 3 explains the highest percentage of variability for Na and Si and, to some minor extent, also for Ti and Al. Presence of Si relates this factor to a silicate source; Na, Ti and Al could be related to different types of silicate minerals.

Factor 4. This factor is characterized by medium-high variability, between 30% and 50%, of Al and K and some transition elements (Ti, V, Mn, Fe and Co). Those elements could identify a geochemical feature of the sediments related to heavy metals-bearing phases and to potassium-aluminium-rich clay minerals.

In **Fig. 31** the contribution of each resolved factor to each lake, normalized to unit sum, is plotted by histograms. From the map, it is evidenced that factor 3 and 4 have a major component in lakes located in the Trentino region, in accordance with a prevalence of volcanic intrusive and metamorphic rocks in the area. In opposition, lakes situated in the Lombardy pre-Alpine zone are subjected to a major impact from factors 1 and 2 in agreement with carbonate rock predominance.

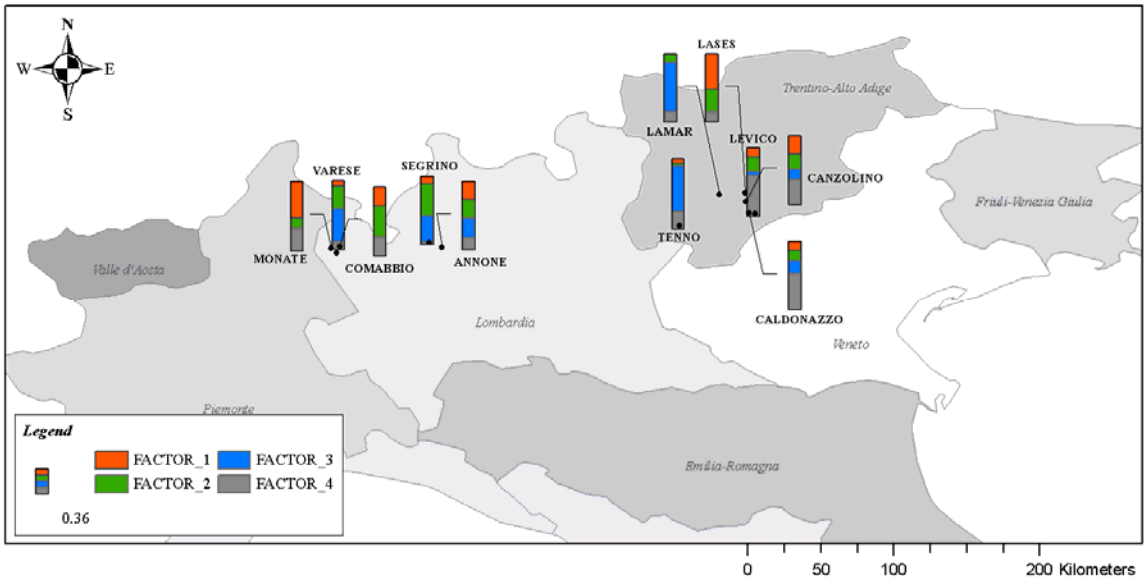


Fig. 31: factor contributions normalized to unit sum.

In order to confirm the mineral composition of the sediments explained by the four interpretable PMF factors, further specific analyses should be made using e.g. X-Ray diffraction technology.

Looking at the **F** explained variation graph, the elements Cr, Ni, Cu and As show NEVF greater than 25%. The reason should be attributed to their medium-high number of BDL observations: 19%, 59%, 28% and 56%, respectively, indicating also the limits of applicability of XRF at these levels. Mg and Mn show NEVF values slightly above 25%, probably due to high element concentration values at some locations.

In addition, also Pb shows a relatively high NEVF value. Examining the Pb concentration plot (**Fig. 32**), high values were observed in a particular lake, making the Pb trend very inhomogeneous. Since PMF analysis could have treated these anomalous values as outliers, in order to better reproduce the Pb trend and to attempt finding hidden information, a new PMF test was made, reducing Pb error estimates by a factor of 2 and operating in the non-robust mode.

A 5-factor solution was determined, where Pb was isolated in a single factor explaining 70% of Pb variability. The remaining four factors have the same characterization of the previous 4-factor solution, with little changes in the explained variation values. The new Pb factor was interpreted as a contamination source. The proximity (about 10 km) of an ancient mining centre for lead, operating until the early 1500s and the presence of a waste matter dump from porphyry mining near the lake subjected to high Pb levels, could support the contamination source hypothesis.

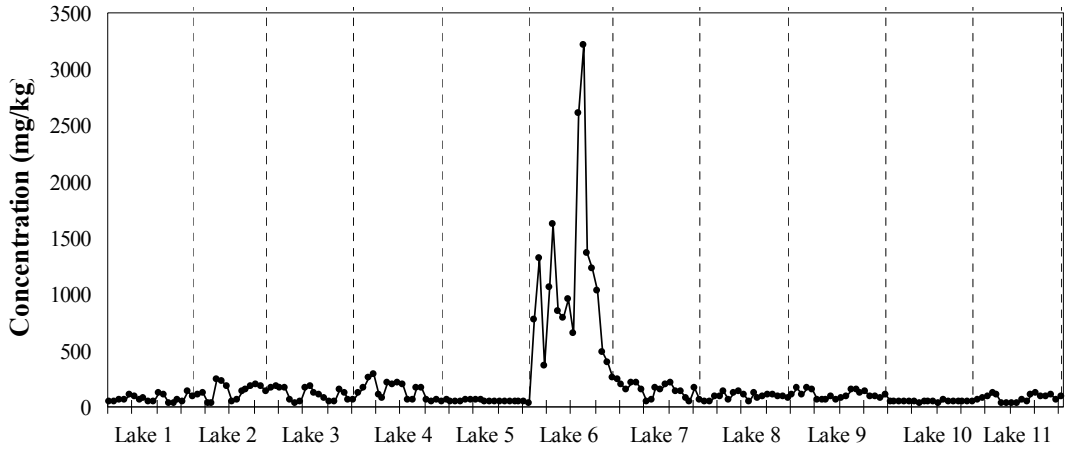


Fig. 32: plot of Pb concentration, expressed in mg/kg, in the examined lakes.

7.4. CA and PCA comparison

In PCA and CA application, Cr, Ni, Cu and As were omitted from the analysis because they show high percentage of BDL (>5%). Moreover outliers were discharged from the data set.

Logarithmic transformation, recommended by Webster (2001) when skewness coefficient is bigger than 1, and z-standardisation procedures were applied to the data-set. R software (R Development Core Team, 2005) was used to perform CA and PCA techniques.

7.4.1. Cluster analysis

Ward agglomerative hierarchic method and Euclidean distance were employed to cluster variables, in order to find groups that show a similar behaviour. In the dendrogram of variables (**Fig. 33**), two main clusters were distinguished, each one split in two sub-clusters.

The first cluster contains Mn, Cl, Zn and Pb, and seems to be connected with a contamination source. However, this cluster could also be due to the grouping of elements that show a high variability (see box-plots in **Fig. 24**).

It is possible that this cluster came from the high order of dispersion of the data within each variable, as some elements exhibit different concentration ranges depending on the lake, as a consequence of the nature of regional geochemical data. The nature of the other clusters did not have a clear interpretation.

Cluster Analysis can also be used to group observations (sampling locations) in order to find homogeneous groups of samples. Dendrogram of location pattern resulted in two main alpine lakes groups: the first cluster represents locations with the highest calcium content, while the second group identify samples composed by a high amount of Al, Si and some other metals.

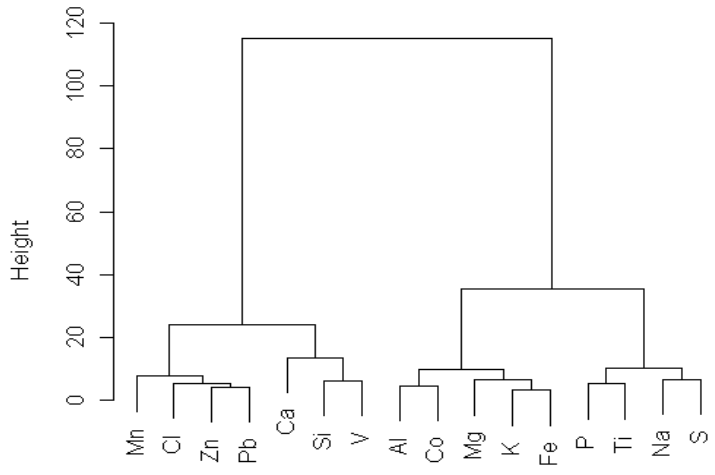


Fig. 33: Dendrogram of Ward agglomerative hierarchic method with Euclidean distance for variables.

7.4.2. Principal component analysis

PCA was performed by the singular value decomposition (SVD) algorithm. Principal components (PCs) with eigenvalues greater than 1 were selected (Kaiser criterion). Eigenvectors of the first three PCs are reported in **Tab. 10**, together with their associated variances.

Variable loadings (**Fig. 34-a**) indicate that PC1 explains 48% of the data variability. Positive loading for Ca were interpreted as a carbonate component, common to factor 2 resulting from PMF analysis. On the other side, negative loadings for Al, K, Ti, V, Fe and, to a lower extent, for Co, Si and Na were related to a silicate and metal-bearing minerals source, as compared to PMF factors 3 and 4.

PC2 accounted for 18% of the total variance and showed negative loadings for S, Zn and Pb (**Fig. 34-b**), suggesting a possible presence of sulphides (sphalerite and galena) and sulphates. PC3, accounting 13% of variance, is dominated by negative Mg loadings, which has no visible relationship with the rest of the elements; this is quite ambiguous as usually Mg is associated both with carbonate or silicate minerals.

Tab. 10: Loadings, variance and cumulative variance for PC1, PC2 and PC3 resulting from PCA analysis.

Variables	*PC 1*	*PC 2*	*PC 3*	*PC 4*
Na	-0.25	0.28	0.17	-0.34
P	-0.07	-0.31	0.39	-0.21
S	-0.06	-0.46	0.30	-0.00
Ti	-0.34	0.08	-0.03	-0.09
Mg	-0.06	-0.07	-0.59	0.15
Al	-0.35	0.10	-0.08	-0.02
K	-0.33	0.09	-0.12	0.07
Fe	-0.34	-0.11	-0.03	0.06
Si	-0.28	0.20	0.22	-0.13
Ca	0.24	-0.33	-0.28	0.10
V	-0.33	-0.09	-0.03	0.06
Co	-0.30	-0.06	-0.30	0.02
Cl	0.13	-0.26	-0.14	-0.67
Mn	-0.13	-0.22	-0.33	-0.47
Zn	-0.23	-0.40	0.06	0.17
Pb	-0.21	-0.37	0.13	0.27
% Variance	48	18	13	5
% Cum.variance	48	66	79	84

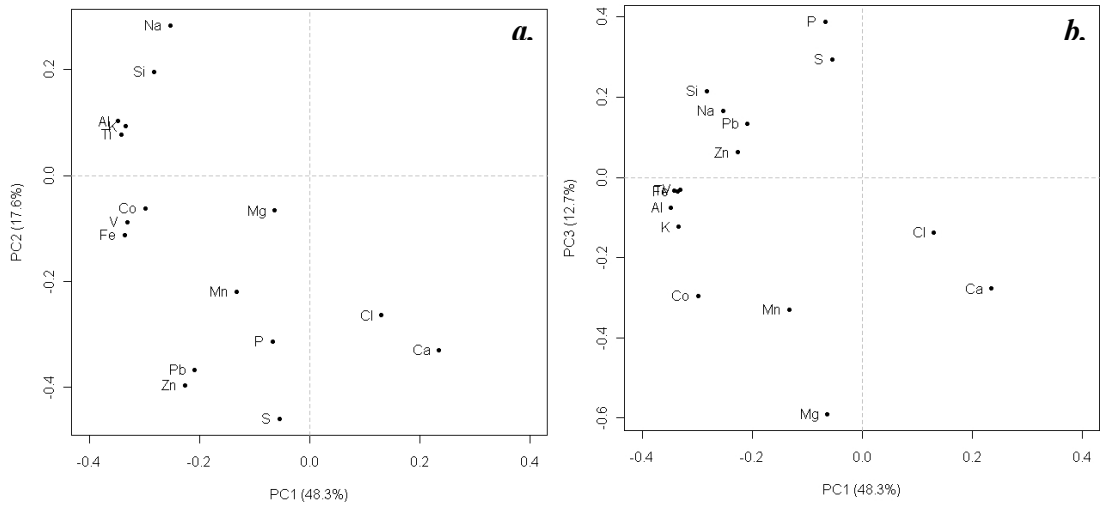


Fig. 34: Plot of PCs extracted from PCA; the amount of the explained variance is indicated in brackets.

7.5. Conclusions

Analysing the results obtained by the three statistical techniques, cluster analysis seems to be the less appropriate approach to handle the data set under examination, characterised by high data variability. In this case, CA should be more appropriate to cluster observations, in order to find groups of samples that show similar features.

Principal component analysis and positive matrix factorization produced similar results. Both techniques identify sources of sulphides and carbonate minerals. Alumino-silicate and metals-bearing minerals components were determined in two different PMF factors, while they were grouped into a single component in PCA. In addition, loadings obtained from PCA showed also negative values making them not directly associated to a real physical meaning.

In conclusion, the positive matrix factorization approach is well adapted to analyse the study data set, with single data uncertainties used to better handle inhomogeneous distributions of variables. Moreover, properly modifying Pb uncertainty estimate, a new factor was resolved, identifying a possible Pb contamination source.

Chapter 8

Application 3 - Danube River

PMF was here applied in a pan-European monitoring exercise to determine how the positive matrix factorization approach adapts in the identification of pollutant sources in a wide area, the Danube river basin.

The Danube is the second longest river in Europe, flowing for 2857 km from the Germany's Black Forest to its delta on the Black Sea. In the past, monitoring programmes were carried out in various parts of its drainage basin, including its tributaries, in order to monitor the micorpollutants level in the river (Literathy and Laszlo, 1995; Sakan *et al*., 2009; Bird *et al*., 2010; Milačič *et al*., 2010).

In 2007, a harmonized monitoring survey, called *Joint Danube Survey* (JDS2) was carried out to investigate the chemical and ecological status of the Danube river basin (ICPDR, 2008). During the JDS2 campaign water, sediments, suspended solids and mussel samples were collected at several representative sampling sites. The various samples were analysed in specific laboratories for different chemical and biological parameters (Woitke et al., 2003).

Bottom sediments play an important role to assess the heavy metals pollution status of a river. In fact, they receive heavy metals from the water column and act as an accumulation reservoir for these contaminants (Literathy and Laszlo, 1995). The main anthropogenic metals discharges in the river basins may come from different type of activities, like industries, mining, agriculture and municipalities (Pizarro et al., 2010; Klaver *et al*., 2007; Santos Bermejo *et al*., 2003). However, also natural processes can affect the river quality, by means of high concentrations of heavy metals influenced by the presence of specific geochemical and mineralogical features (Keshav Krishna et al., 2011).

In the case under study, being that the mineralogy of the Danube is very complex (Yiğiterhan and Murray, 2008) due to the heterogeneity of rock types present along its course, attention must be paid to discriminate the anthropogenic impact from the natural background values of heavy metals sediment content (Devesa-Rey *et al*., 2009).

Usually, the enrichment factors (EF) method, with the use of an appropriate normalising element not affected by anthropogenic sources, and a geochemical background, is applied to determine the anthropogenic contribution (Devesa-Rey *et al*., 2009; Woitke *et al*., 2003). However,

reference values for sediments are not always available and comparison with average crustal values may be not appropriate if the studied area is very heterogeneous.
Here, the PMF approach was used to determine the natural vs. anthropogenic origin of heavy metals. Moreover, the spatial distribution of resulting sources was helpful to determine the role of Danube tributaries as potential sources of pollution.

8.1. Site characterization

The Danube River catchment covers a very wide area (817.000 km^2), flowing through nine countries (Austria, Bulgaria, Croatia, Germany, Hungary, Serbia, Slovakia, Romania and Ukraine). The rock types outcropping along the river basin are very different both for lithologic composition and for age (Yiğiterhan and Murray, 2008). They includes igneous and metamorphic Precambrian and Paleozoic rocks of the Bohemian Massif, Mesozoic carbonate sediments, young orogenic belts of the Alps and Western Carpathians, and Cenozoic sediments of the Alpine molasse, only if one considers the section between the source and Hungary. In the Hungarian plain, the river flows over Olocene alluvium, made by sediments different both for grain size, from gravels to muds, and for chemical composition. The western part of the Southern Carpathians, the Banat Mountains and the mountains of eastern Serbia, at the Iron Gate, are split apart by the gap valley of the Danube. They represent the last reliefs, mostly made by silicate rocks (igneous and metamorphic) that the river meets before its flow in the Romanian plain (Walachia), this last characterized by Pleistocene loess sediments.
Drainage basins of most tributaries are dominated by the same lithologies affecting the Danube course, probably with a greater contribution from sedimentary lithologies. The tributaries involved in the sampling campaign were the following: Iskar, Timok, Velika Morava, Ipoly, Vah, Sava, Moson Arm, Sio, Jantra, Tisza, Rackeve-Soroksar Arm, Hron, Arges, Sulina arm (old Danube), Bystroe canal, Russenski Lom, Szentendre Arm, Siret, Prut, Olt and Inn.
For this reason the catchment area was divided in nine different reaches by Vogel and Pall (2002), listed in **Tab. 11**, which were selected basing on both the geo-morphological classification and the anthropogenic impact.
During the JDS2 campaign, a total of 148 bottom sediment samples were collected from both Danube River and its tributaries. Sampling sites were grouped according to the following categories: Danube River (110 sediments), tributary at confluence (23 sediments) and tributary (15 sediments).

Tab. 11: Nine geo-morphological reaches of the Danube River basin, from Vogel and Pall, 2002.

Reach	*Characteristic*	*River km*
1	Alpine river character, anthropogenic impact by hydroelectric power plants	2581 – 2225
2	Alpine river character, anthropogenic impact by hydroelectric power plants.	2225 – 1880
3	Anthropogenic impact by the construction of Gabcikovo Dam	1880 – 1816
4	Starting development from alpine to lowland river, the Danube passes the Hungarian Highlands.	1818 – 1659
5	Lowland river; the Danube passes the Hungarian Lowlands; anthropogenic impact by significant emissions of untreated wastewater at Budapest.	1659 - 1202
6	Lowland river; the Danube breaks through the Carpatian and the Balkan Mountains; anthropogenic impact by damming effect on Iron Gate hydroelectric power plant and significant emission input of untreated wastewaters al Belgrade	1202 – 943
7	Lowland river; the Danube flow through the Walachian Lowlands (Aeolian sediments and loess); steep sediments walls (up to 150 m) characterize the Bulgarian river bank.	943 – 537
8	Lowland river; alluvial islands between two Danube arms.	537 – 132
9	The Danube splits into three Delta arms; characteristic wetland and estuary ecosystem; slopes decrease to 0,01‰	132 – 12

Danube sediment samples were collected from both left and right benches of the rivers while, for tributaries only, a single mixed sediment sample was taken (ICPDR, 2008).

A map of the Danube catchment area, showing the sampling site locations, is given in **Fig. 35**.

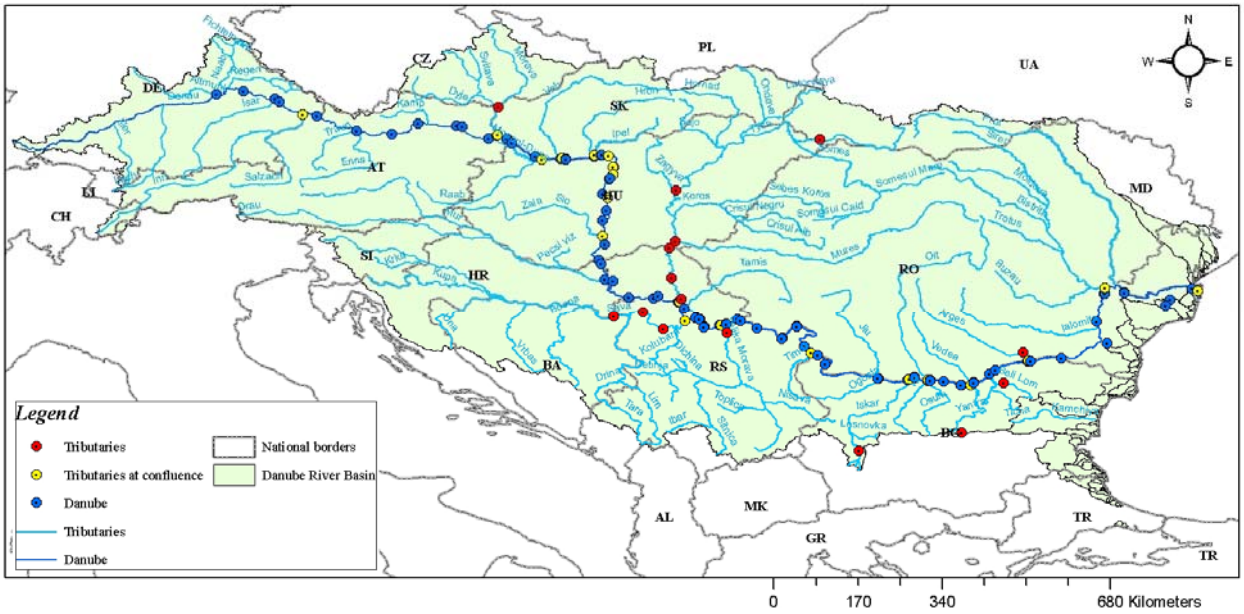


Fig. 35: map of the Danube catchment area. Sampling locations for the Danube River and its tributaries are shown.

8.2. Data set description

Before each analytical measurement, sediment samples were dried in an oven for 24 hours, whose air temperature did not exceed 40 ºC. Then, samples were milled for about 5 minutes, using a planetary mill provided by an agate-zirconia milling vessel.

Major and minor elements, and heavy metals were detected by means of a wavelength dispersive X-ray fluorescence (WD-XRF) spectrometer, Bruker AXS® SRS-3400 device. The following elements were measured: Al, As, Ca, Cd, Cl, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, P, Pb, S, Si, Ti, V and Zn.

Prior to each sediment analysis, about 2 g of sample were pressed into pellets, using a hydraulic press operating at a pressure of 20t/cm^2, applied for 20 seconds. The instrument was calibrated using the following certificate reference material for soils and sediments: BCR-141, BCR-141R, BCR-142, BCR-142R, BCR-143, BCR-143R, BCR-144, BCR-144R, BCR-145, BCR-145R, BCR-146, BCR-146R, BCR-277, BCR-280, BCR-320, CAnMET-SO1, CANMET-SO2, CANMET-SO3, CANMET-SO4, NIST-SRM-2704, NIST-SRM-2709, NIST-SRM-2710, NIST-SRM-2711, IAEA-SOIL-7. Fixed alpha correction, computed by empirical regression method, was applied to correct matrix effects. For energies range beyond the Fe Kα line, the matrix correction was applied using the Rh Kα Compton scattered tube line as an internal standard. All measurements were run under repeatability conditions.

Mercury was analysed in dried and milled bottom sediments. Cold vapour-atomic adsorption (CV-AAS) technique was employed by means of the Advanced Mercury Analyser (AMA-254, Leco) instrument. A mercury stock standard solution (Carlo Erba) with a nominal mercury concentration of 1 mg/mL was used to prepare calibration standard solutions, by stepwise dilutions. Calibration curves were tested using the following certificate reference materials: CRM: BCR-141R, BRC-143R, RTH-953.

8.3. Descriptive statistic

In the study data set, only the variable Hg contains missing values, comprising 11% of all samples. Missing values were replaced with the mean value for mercury concentration. Below-detection-limit (BDL) values were detected in a percentage less than 2% for S, Ni and As, and to the extent of 37% for Cd. Since left censored values were known, they were used instead of replacing them with the more common used formula BDL/2. A descriptive statistic was carried

out on the two separate sediments sub-sets: Danube river sites and Danube tributary sites (including both tributaries and tributaries at confluence), in order to underline some possible differences between them.

Measured element concentrations and descriptive statistic for both the Danube and its tributaries sediments is given in **Tab. 12** and **Tab. 13**, respectively.

Comparison boxplots in logarithmic scale for the two groups of sediment are shown in **Fig. 36**. Concentrations were expressed in wt% (weight percentage) except for mercury, which is expressed in μg/g.

Tab. 12: Descriptive statistics of the elements concentration for the Danube River sampling sites.

DANUBE	*Min*	*Max*	*Average*	*Median*	*1st quartile*	*3rd quartile*	*Std.dev.*	*CV (%)*	*Skewness*
Hg (mg/kg)	0.04	1.33	0.30	0.28	0.16	0.37	0.19	65%	2.02
Al (%)	5.05	9.86	7.60	7.72	7.09	8.24	0.89	12%	-0.64
As (mg/kg)	35.0	104	60.6	61.0	53.0	68.0	10.9	18%	0.50
Ca (%)	1.75	14.7	6.89	6.81	4.64	8.89	2.52	37%	0.31
Cd (mg/kg)	< 8	15.0	9.39	9.00	8.00	11.0	2.09	22%	0.27
Cl (%)	0.01	0.03	0.01	0.01	0.01	0.01	0.003	33%	2.32
Co (mg/kg)	9.00	30.0	19.0	19.0	16.0	22.0	4.25	22%	-0.05
Cr (mg/kg)	74.0	208	146	140	121	172	31.0	21%	-0.14
Cu (mg/kg)	43.0	416	85.1	75.0	67.0	89.0	42.4	50%	4.91
Fe (%)	2.42	5.70	4.15	4.19	3.85	4.64	0.65	16%	-0.52
K (%)	1.24	2.48	1.97	1.96	1.88	2.07	0.22	11%	-0.49
Mg (%)	1.27	3.62	2.27	2.07	1.70	2.78	0.64	28%	0.47
Mn (mg/kg)	621	1794	1059	1017	899	1215	235	22%	0.72
Na (%)	0.33	0.93	0.71	0.71	0.64	0.78	0.11	16%	-0.98
Ni (mg/kg)	48.0	195	91.7	86.5	74.0	107	26.4	29%	0.92
P (%)	0.07	0.24	0.11	0.11	0.10	0.12	0.02	19%	2.34
Pb (mg/kg)	39.0	181	72.8	64.0	56.0	84.8	23.7	33%	1.98
S (%)	< 0.01	0.48	0.10	0.10	0.07	0.13	0.06	59%	3.41
Si (%)	16.5	26.6	22.8	23.7	20.4	25.0	2.4	11%	-0.26
Ti (%)	0.26	0.63	0.47	0.46	0.40	0.54	0.08	18%	-0.14
V (mg/kg)	60.0	153	108	111	95.0	121	18.4	17%	-0.43
Zn (mg/kg)	119	575	233	192	160	271	97.6	42%	1.19

Tab. 13: Descriptive statistics of the elements concentration for Tributaries and Tributaries at confluence of Danube River sampling sites.

TRIBUTARIES	*Min*	*Max*	*Average*	*Median*	*1st quartile*	*3rd quartile*	*Std.dev.*	*CV (%)*	*Skewness*
Hg (mg/kg)	0.01	1.42	0.30	0.26	0.11	0.40	0.27	91%	2.48
Al (%)	4.54	9.18	7.80	7.79	6.98	8.62	0.97	12%	-0.88
As (mg/kg)	< 5	272	65.8	60.0	48.3	74.8	39.8	60%	3.91
Ca (%)	1.04	8.98	4.71	4.41	2.86	6.45	2.26	48%	0.35
Cd (mg/kg)	< 8	21.0	9.79	9.50	7.00	12.00	3.71	38%	0.72
Cl (%)	0.01	0.05	0.01	0.01	0.01	0.01	0.01	55%	3.81
Co (mg/kg)	8.00	67.0	22.6	20.0	15.0	27.0	9.95	44%	2.42
Cr (mg/kg)	65	283	159	158	128	176	46.4	29%	0.46
Cu (mg/kg)	28.0	13666	459.2	70.5	58.3	101	2203	480%	6.14
Fe (%)	2.09	10.07	4.46	4.36	3.66	5.06	1.22	27%	2.50
K (%)	1.25	2.49	2.02	1.99	1.85	2.19	0.26	13%	-0.44
Mg (%)	0.84	3.12	1.66	1.51	1.34	1.85	0.56	34%	1.06
Mn (mg/kg)	530	2472	1301	1193	994	1546	456	35%	0.98
Na (%)	0.50	1.14	0.69	0.68	0.61	0.77	0.12	17%	1.48
Ni (mg/kg)	< 5	254	103	86.5	65.8	103	61.0	59%	1.24
P (%)	0.07	0.33	0.12	0.11	0.10	0.13	0.04	37%	2.83
Pb (mg/kg)	41.0	393	88.9	61.5	51.0	98.0	64.7	73%	3.18
S (%)	< 0.01	1.30	0.16	0.10	0.07	0.18	0.21	131%	4.57
Si (%)	20.1	28.8	24.9	25.2	23.6	26.5	2.28	9%	-0.26
Ti (%)	0.32	0.68	0.50	0.51	0.46	0.54	0.08	15%	-0.29
V (mg/kg)	52.0	166	116	117	97.8	132	23.0	20%	-0.40
Zn (mg/kg)	102	1070	300	241	148	420	207	69%	1.92

From boxplots of **Fig. 36**, it is observed that mean elemental concentrations in the Danube River and its tributaries do not vary significantly. However, a wider spread in the majority of elemental concentration data is observed for the tributaries data set. This reflects a higher degree of variation in the chemical composition of sediments, in the contest of different sub-basins areas for tributaries. This tendency was also observed in the first JDS campaign, JDS1 (Woitke *et al.*, 2003).

In the Danube data set, the distribution of Hg, Cl, Cu, P, Pb, S and Zn is high positively skewed (**Tab. 12**), with skewness coefficient >1, indicating possible hotspots which could have both a natural or anthropogenic origin. For tributaries, data distributions exhibit a high skewness also for As, Co, Fe, Mg, Na and Ni (**Tab. 13**).

It is however to consider that the number of sampling location is lower for the tributaries data set (38 samples for tributaries and 110 for Danube river), making the tributaries statistic less representative.

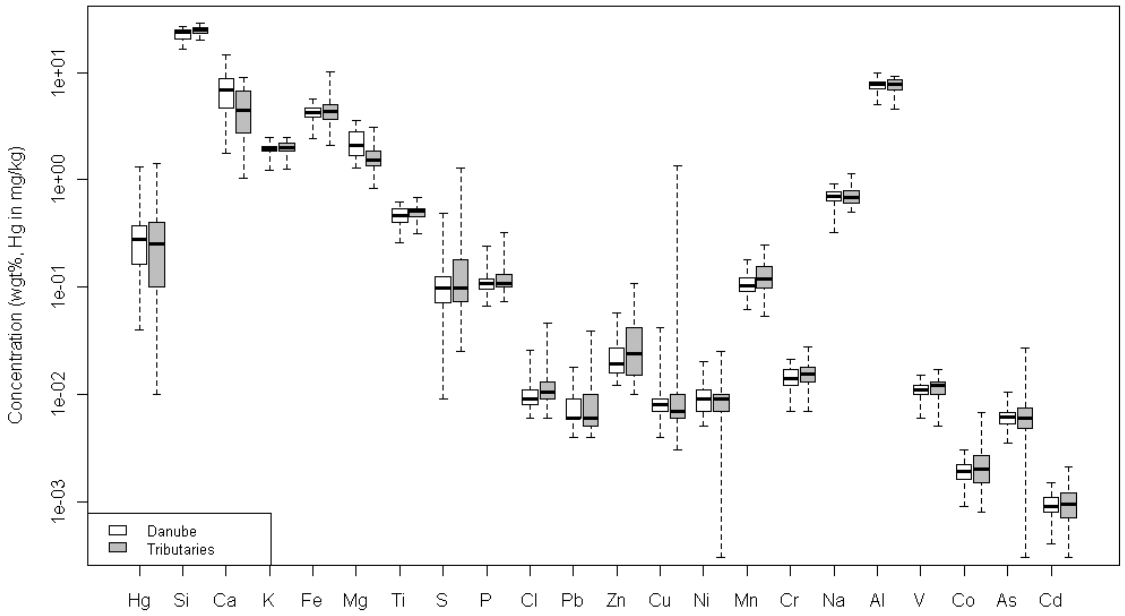


Fig. 36: Boxplot showing the variation of measured element concentrations: median, 1st and 3rd quartiles and whiskers (lowest and highest values). White boxes are for Danube River (right) and grey boxes for Tributaries (left).

8.4. Positive matrix factorization

The PMF analysis was performed in the *robust mode* using an outlier distance equal to 4. From 2 to 8-factor solutions were investigated, together with the *FPEAK* parameter ranging between -1 and +1, with a 0.1 incremental step.

The error estimate data matrix was computed by means of the error model EM=-14, directly implemented into the resolving algorithm:

$$s_{ij} = t_j + \nu_j \cdot \max\left(\left|x_{ij}\right|, \left|y_{ij}\right|\right)$$

The formula includes both the contribution coming from the original input data x_{ij} or the fitted values y_{ij}. The multiplier factor ν_j represent the relative uncertainty in the data measurements, while the t_j coefficient is the computed detection limit for each measured element. Values for t_j and ν_j parameters, used in this study, are given in **Tab. 14**.

For mercury only, the relative uncertainty parameter (t_j value) was increased by a 2 factor, in order to take into account the high percentage (11%) of missing values. Cadmium, which shows a high percentage of BDL, was not down-weighted due to its elevated relative uncertainty compared to the other elements.

Tab. 14: t_j and v_j values used in the EM=-14 error model equation.

Element	*Hg*	*Al*	*As*	*Ca*	*Cd*	*Cl*	*Co*	*Cr*	*Cu*	*Fe*	*K*
t_j	0.005	0.003	5	0.003	8	0.005	2	3	5	0.001	0.0005
v_j	0.2	0.05	0.1	0.05	0.5	0.20	0.1	0.1	0.1	0.05	0.05

Element	*Mg*	*Mn*	*Na*	*Ni*	*P*	*Pb*	*S*	*Si*	*Ti*	*V*	*Zn*
t_j	0.0015	70	0.003	5	0.0003	5	0.010	0.09	0.0004	2	5
v_j	0.15	0.05	0.05	0.1	0.05	0.1	0.01	0.05	0.05	0.05	0.1

Aiming at the determination of the optimum solution, Q and MaxRotMat values were plotted against the examined number of factors (**Fig. 37**). Moreover IM and IS parameters were investigated (**Fig. 38**).

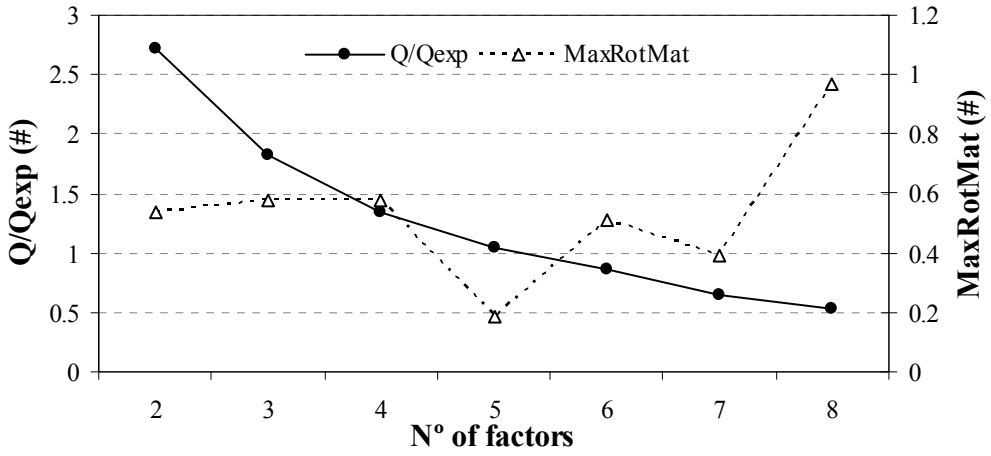


Fig. 37: Q vs. Q expected (left) and RotMat (right) parameters for each number of factors examined.

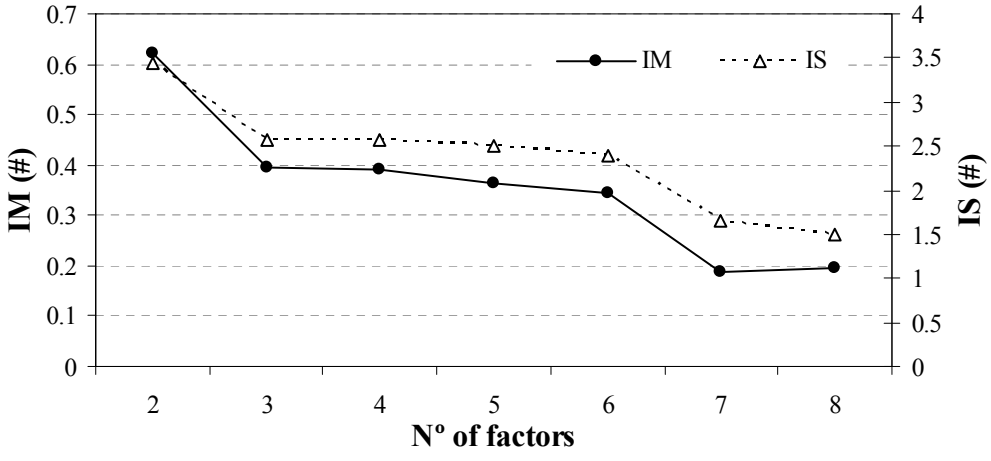


Fig. 38: IM and IS parameters values for each examined number of factors.

Solutions with more than 5 factors were not considered because they show a Q value lower than the expected Q. Moreover the 2-factor solution was omitted since shows high IM and IS values.

Looking at the not explained variations for the selected solutions (3, 4 and 5 resolved factors) in **Tab. 15**, it is clear that no significant changes in the NEVF appear for any elements. Passing from 3 to 4 factors, the most significant change in NEVF is for Si which, in the 4-factors solution, is classified in two distinct factors. However, one of the two Si-factors, which shows low explained variation (about 30%), has not a clear interpretation.

Tab. 15: Not explained variations of F for solutions with 3, 4 and 5 factors extracted.

	Number of factors		
	3	4	5
Hg	25%	25%	26%
Si	8%	3%	3%
Ca	5%	5%	2%
K	5%	4%	4%
Fe	5%	3%	2%
Mg	6%	4%	2%
Ti	4%	3%	3%
S	20%	21%	20%
P	6%	4%	5%
Cl	19%	18%	18%
Pb	13%	11%	10%

	Number of factors		
	3	4	5
Zn	17%	11%	8%
Cu	17%	15%	14%
Ni	14%	14%	13%
Mn	12%	12%	12%
Cr	8%	8%	7%
Na	8%	7%	2%
Al	2%	2%	2%
V	4%	3%	2%
Co	10%	8%	8%
As	12%	11%	11%
Cd	13%	12%	12%

The same conclusions could be also applied for the 5-factor solution, in which NEVF show a significant decrease for Ca and Na. Calcium was explained by two distinct factors, one of them with lower explained variations.

Rotations did not alter the factor interpretation, showing very low differences in variables EVFs. For these reasons, the 3-factor solution was chosen as the most representative. The solution with FPEAK = 0.2 was selected. Rotated and central Q values for the 3-factor solution differ for less than 1% for all the selected rotations (**Fig. 39**). The MaxRotMat parameter (**Fig. 40**) permits to exclude rotations closed to the central solution, in particular with *FPEAK* between -0.2 and 0.1. The IM and IS parameters decrease on both sides of the central solution.

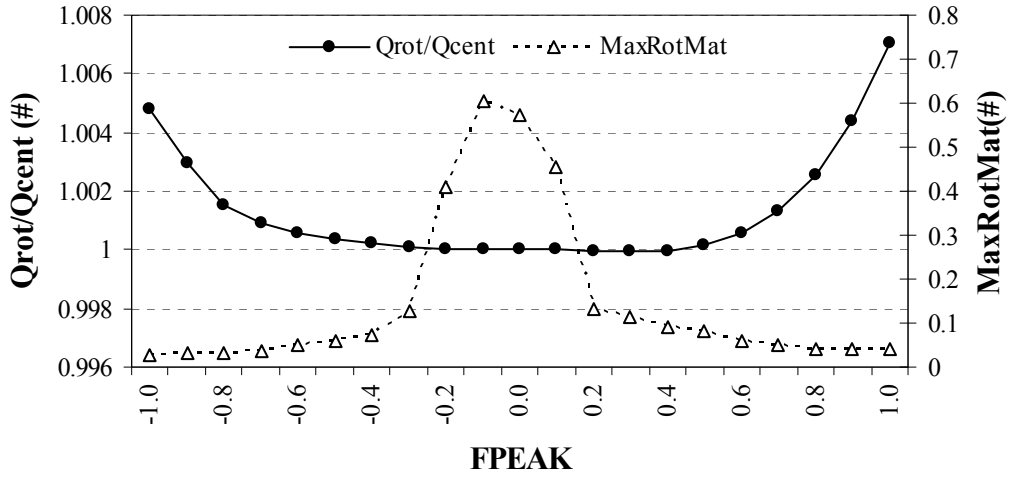


Fig. 39: Q for rotations vs. Q for central solution (left) and RotMat (right) parameters varying the FPEAK value.

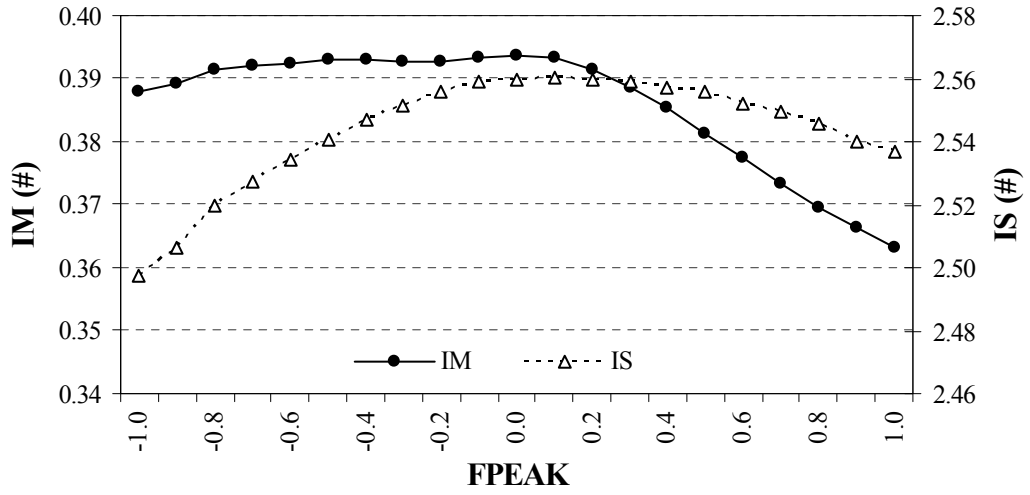


Fig. 40: IM and IS parameters for different FPEAK values tested.

Finally, the *FPEAK* = 0.2 was chosen basing on the fact that the IM parameter, which shows the higher variation (8%), start to decrease with a higher slope and this value. However, EVFs did not change significantly varying the rotational degree.

Explained variations for the 3-resolved factors are shown in **Fig. 41**. Moreover, **G** matrix elements (score matrix), representing the contribution of each resolved factor to the sampling sites, were used better understand the source interpretation in relation to their geographical distribution. In the score map representation, reported in Fig. 42, the sampling sites locations were plotted using graduated symbols (differently sized points) classified in 5 categories, using the natural breaks (Jenks).

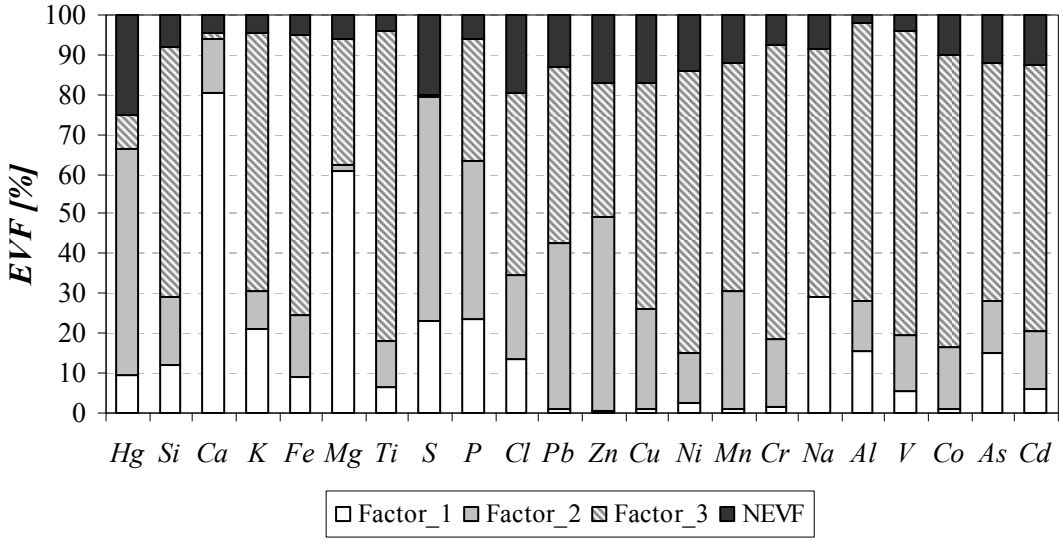


Fig. 41: Explained variations of F matrix for the 5-factor solution with FPEAK=0.2.

Factor 1

The EVF bar-plot for factor 1 shows high values for Ca and Mg, with a contribution of 80% and 61%, respectively. Being the factor uniquely explained by these two elements, which are mainly linked to carbonates, a carbonaceous source was suggested. Observing the source distribution illustrated in **Fig. 42-a**, this factor appears to be more correlated with the Upper and Middle part of the Danube River. In particular, referring to the catchment areas of the river (**Tab. 11**), the carbonaceous source is more representative for the reaches identifying an Alpine stream (reaches 1, 2 and 3), and a lowland river (reaches 4 and 5). This is in agreement with the break of the Danube through the Carpathian and the Balkan mountains which start from reach 6.

The factor explanation also agrees with a predominance of carbonates in the upper drainage basin, due to the Mesozoic carbonate complexes of the Alps (Pawellek *et al*., 2002). Dissolved carbonate could in fact lead to an increasing concentration of Ca and Mg in the sediments. In Pawellek *et al*. (2002) it was also found that the silicon dioxide concentration in the Upper part of the Danube is below typical natural values found for the major world rivers.

Moreover, scatter plots of Al vs. Ca and Mg show a negative correlation, confirming the source identification; a positive correlation was instead observed for Ca and Mg (r^2=0.73).

Factor 2

Factor 2 is characterized by the variation of S (56%) and P (40%), the metals Hg (57%), Zn (49%), Pb (42%) and, to a lesser extent, Mn (30%) and Cu (25%). The common association of these metals with various forms of environmental pollution suggests an anthropogenic source for

their origin (Bird *et al*., 2010, Milačič *et al*., 2010). Since the study area is very extended and includes different geological territories as well as urbanized districts, the interpretation of factor 2 could be improved by the use of the factor scores map (**Fig. 42-b**). Highest **G** values are mainly localizes in three different areas: (i) the tributaries, represented by white squares in **Fig. 42-b**; (ii) the Middle part of the Danube, located in reaches 5 and 6; and (iii) the very Upper part of the river situated in reach 1. Moreover some hotspots might be identified along the Danube flow.

The majority of the tributaries are influenced by this anthropogenic source, with highest **G** values found for Iskar and Velika Morava rivers. Their metals content in sediments could be influenced by mining activity in the catchment area, in which enrichment of heavy metals were found (Bird *et al*., 2010).

In Sava River, the biggest tributary of the Danube, elevated concentrations of mercury were found in a previous study (Milačič *et al*., 2010), probably in association with oil refinery activities and chemical industry. Tisza river was in the past contaminated by industrial accidents resulting in cyanide and heavy metals spill at Baia Bare and Baia Borsa, respectively (Sakan *et al*., 2009). Morava river was instead subjected to agriculture and municipal waste water discharges (Gashi *et al*., 2011).

A common pollutant source to all the listed tributaries, which could be identify by factor 2, might be due to uncontrolled discharge from municipalities (UNECE, 2007), characterized by heavy metals loads as well as phosphorus and sulphur content (Hoffman *et al*., 2010, Sheng *et al*., 2011).

High score for factor 2 are also localized in reaches 5 and 6 of the Danube catchment, where the river flows through the Serbia region and the confining countries, Croatia and Romania. In these reaches a strong anthropogenic impact is mainly caused by the emission of untreated wastewater in the Budapest and Belgrade areas, as well as by dumming effect (Vogel and Pall, 2002), which could explain the association of heavy metals, P and S to this factor.

Moreover, in the Serbia region, factor 2 may also be related to the pollution disaster caused by the Kosovo conflict. The bombing of industrial sites, in particular burnings of oil refineries and oil depots, were the origin of a general contamination of air, water and land, with a consequent trans-boundary effect (Melas *et al*., 2000; Relić *et al*., 2005).

An anthropogenic impact was also evidenced in reach 1, located in Germany. This information could reveal the impact caused by the presence of the hydroelectric power plant in Geisling (ICPDR, 2005).

Finally, some hotspots for factor 2 may be identified along the Danube path:

- The majority of the hotspots are located at the confluence of the Danube tributaries, in particular in Timok, Iskar, Ipoly, Vah and Moson. Timok and Iskar, which are affected by mining contamination in the Bulgaria region (Bird *et al*., 2010). Moreover, exploitation of mines and heavy metal industry in Serbia contributes to the heavy metals contamination in Timok (Paunović *et al*., 2008). In the catchment of Vah tributary, strong mercury pollution was found (Woitke *et al*., 2003), while Moson river was affected by high untreated wastewaters discharges released from a municipality (Kirschner *et al*., 2009).
- Two other hotspots are located in proximity of the Oltenita (downstream Arges tributary) and Baja cities. These sites are probably affected by the pollution originated from the cities runoff (Bostan *et al*., 2000).

Factor 3

This factor is characterized by high variations, between 50% and 80%, for Al, Fe, K, Na, Si and Ti, and for the heavy metals As, Cd, Co, Cr, Cu, Mn, Ni and V. Moreover, factor 3 is also determined, to a minor extent (30-50%), by Cl, Mg, P, Pb and Zn variation. The connection between heavy metals and Si, Al and Fe content of sediments, suggest a background component for this source, originating from alumino-silicates and oxide phases.

A significant influence of trace elements content in natural background was also found in the past (Literathy and Laszlo, 1995). This type of background composition was also characterized by Sakan *et al*. (2010), in the Serbian catchment of Danube.

In **Fig. 42-c,** high factor 3 scores were observed in the Lower Danube (reaches 6, 7, 8 and 9), indicating a predominance of metals bounded to alumino-silicates and oxides component in this territories. This background characterization is opposed to factor 1, which dominates the Upper and Middle Danube. This is in agreement with Woitke *et al*. (2003) study, which found an increase in heavy metals concentration from the Iron Gate reservoir (reach 6) to the Danube Delta (reach 9) in the JSD sediments.

Fig. a)

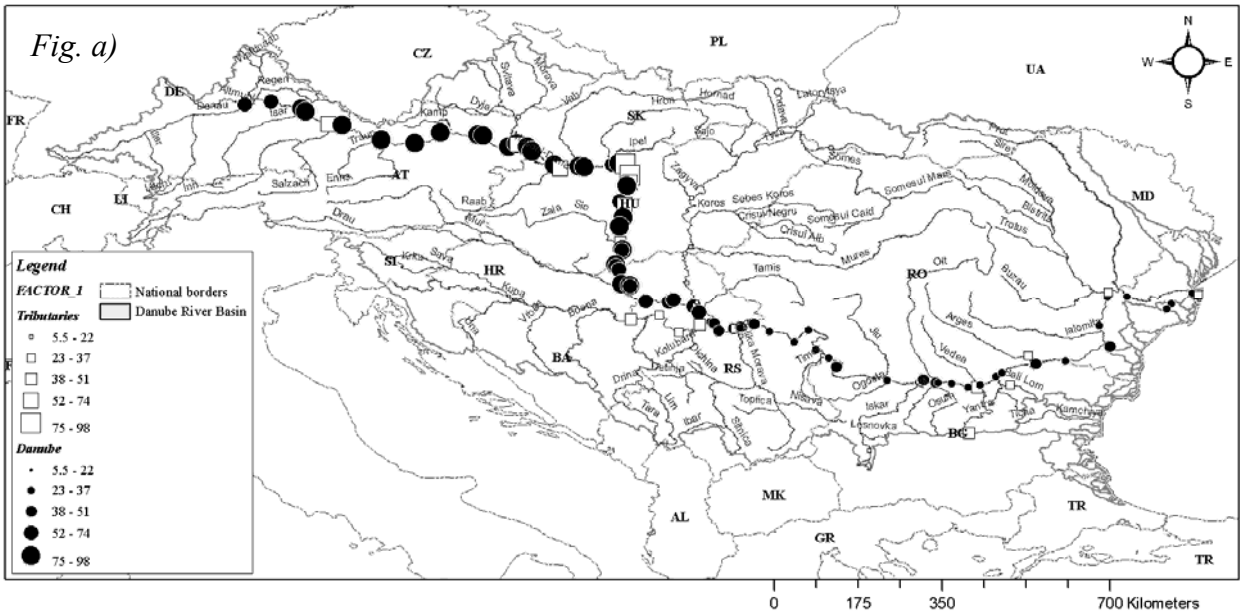
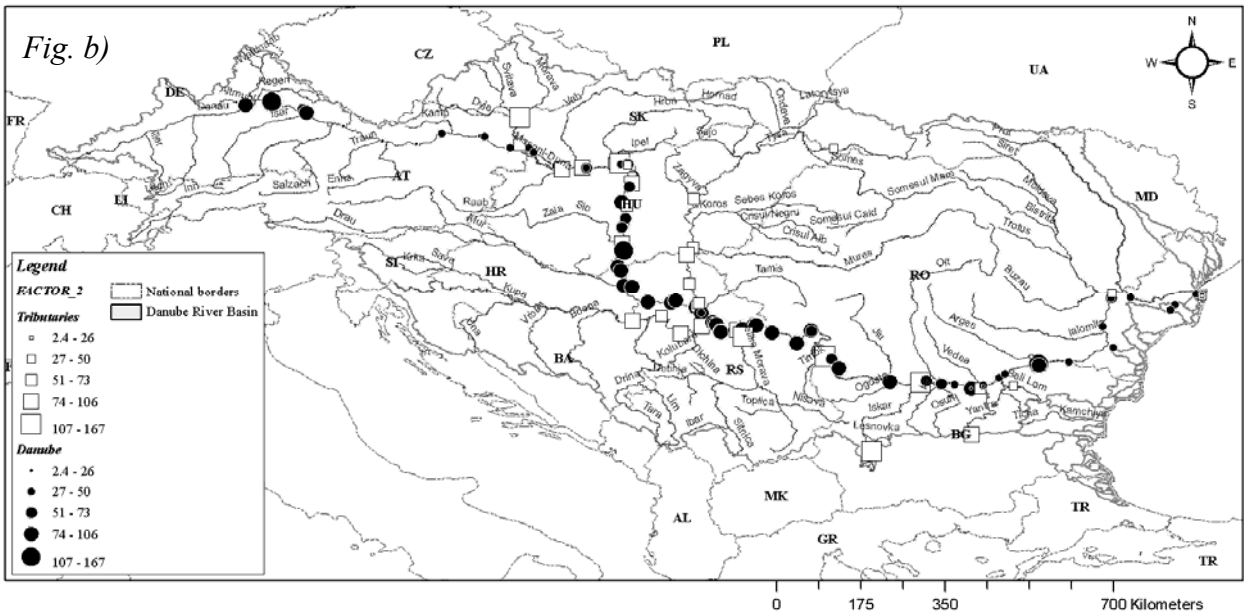

Legend
FACTOR_1
National borders
Danube River Basin
Tributaries
5.5 - 22
23 - 37
38 - 51
52 - 74
75 - 98
Danube
5.5 - 22
23 - 37
38 - 51
52 - 74
75 - 98
0
175
350
700 Kilometers

Fig. b)


Legend
FACTOR_2
National borders
Danube River Basin
Tributaries
2.4 - 26
27 - 50
51 - 73
74 - 106
107 - 167
Danube
2.4 - 26
27 - 50
51 - 73
74 - 106
107 - 167
0
175
350
700 Kilometers

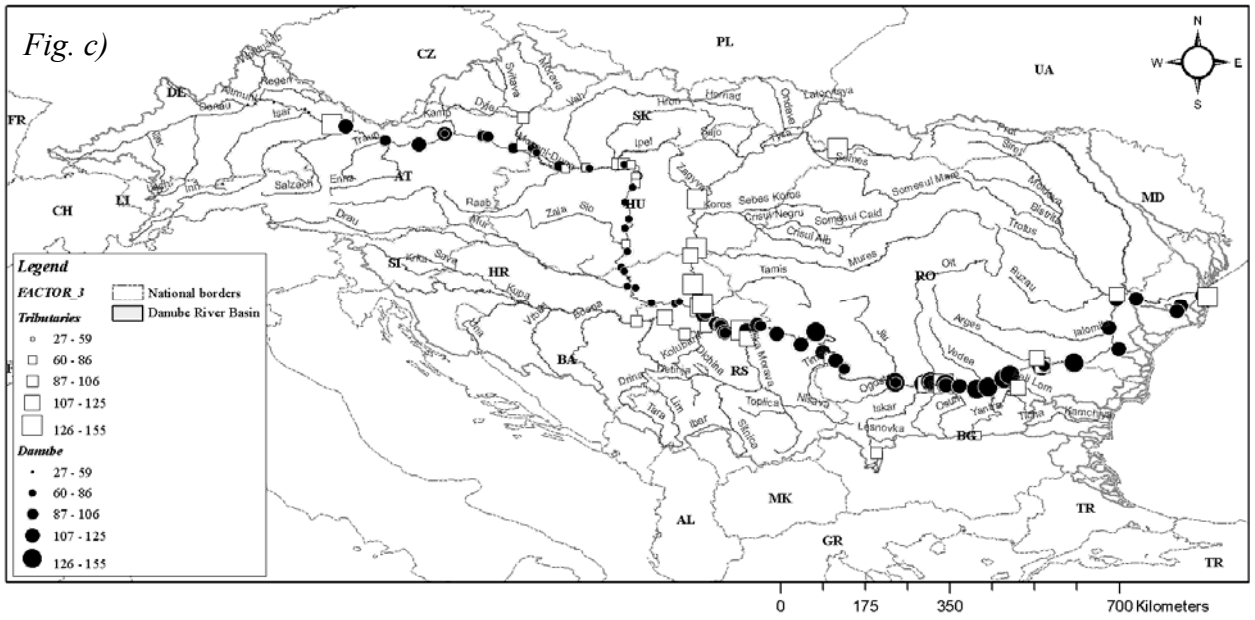


Fig. 42: Factor scores maps of Danube River catchment area. G matrix values were plotted using graduated symbols. Black circles identify Danube River location; white squares represent tributary and tributary at confluence locations.

In conclusion, PMF application identified one anthropogenic factor, which could be connected to different anthropogenic activities depending on the location site along the Danube River: municipal and industrial discharge, and mining activity. Examining their scores, we found a higher impact both in reaches 5 and 6 along the Danube course in Hungary, and in the majority of tributaries and tributaries at confluence. This important information highlights the influence of Danube tributaries.

In order to better understand the role of tributaries, the PMF was further applied on two sub-sets separately. The first data set was determined by the Danube River sites, and the second being composed by tributaries and tributaries at confluence locations.

The application of the model to the Danube data set did not reveal significant changes. Three factors were obtained. Solutions with more than three factors were rejected because the computed Q value was lower than expected Q (**Fig. 43**). The resolved sources could be identified similarly to the sources resolved considering the whole data set (Danube plus tributaries). Only minor variations were detected in the EVF values.

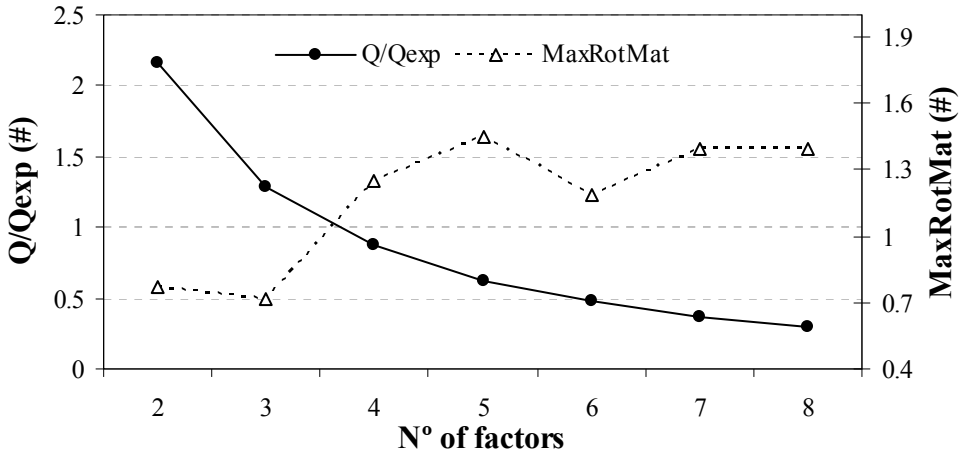


Fig. 43: Q vs. Q expected (left) and RotMat (right) parameters for each number of factors examined. Danube data-set only.

PMF was then applied to the 38 tributaries samples. Considering Q values, solutions with more than 6 factors were rejected. Examining IM and IS (**Fig. 44**) from 3 to 6 factors were further studied to determine the optimal solution. The 4-factor solution was chosen as the most representative, with *FPEAK*=-0.4.

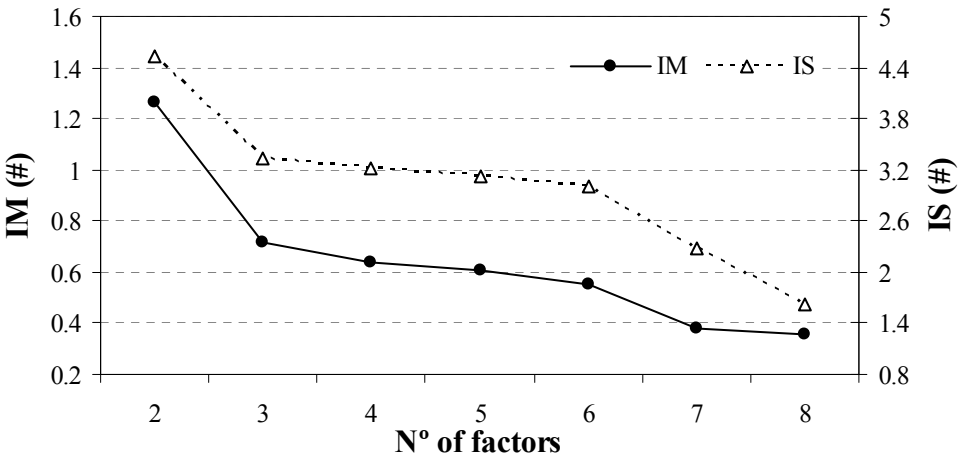


Fig. 44: IM and IS parameters values for each examined number of factors. Only tributaries and tributaries at confluence locations.

Explained variations for the tributaries data set are shown in **Fig. 45**. Basing on EVF interpretation, factor 1 and 4 were analogous to the two natural background components previously obtained considering the whole data set. In particular, factor 1 is representative for a carbonates source, while factor 4 is characterizes by metals bounded to alumino-silicates and oxides phases.

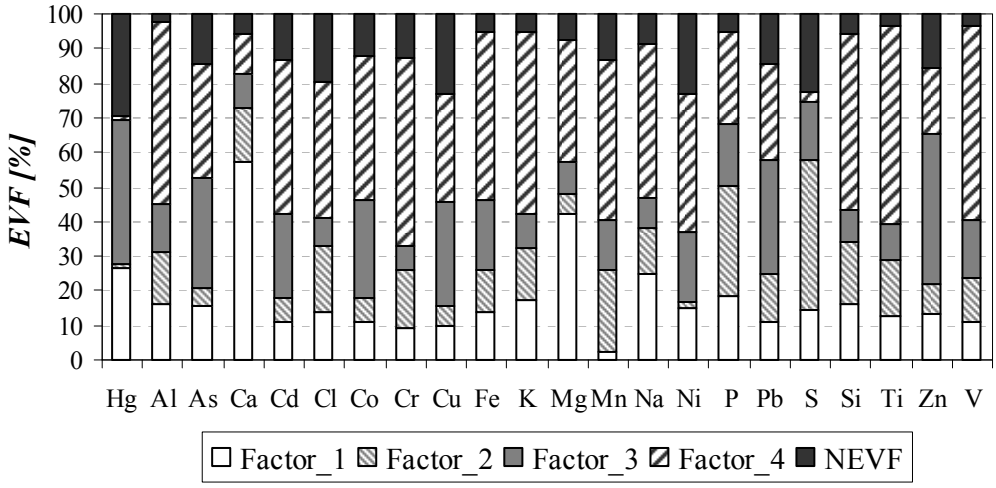


Fig. 45: Explained Variation of F for tributaries sites only, computed by PMF. Also Not Explained Variations were reported.

Factor 3 is characterized by As, Co, Cu, Hg, Pb and Zn and, to a minor extent, by Cd variations. This heavy metals association suggests an anthropogenic origin for their concentration in sediments. However, opposite to the previously determined pollution source found for the Danube data set, in the tributaries case As, Cd and Co compete in the explanation of the anthropogenic factor, while phosphorus and sulphur are missing. This could indicate an anthropogenic component more connected with mining activity and industrial facilities discharge, rather than uncontrolled municipal discharge. The source is found to have a higher impact in the Timok, Iskar, Tisa, Velika Morava and Sava tributaries (reaches 5, 6 and 7), confirmed by the influence of mining industry, including solid waste disposal, on the listed tributaries pollution (Bird et al, 2010; Sakan et al, 2009; UNECE, 2007).

Finally, factor 2 is characterized by S and P, and to a minor extent, by Mn. Their association suggests that a nutrient pollution source stems from agriculture, mainly due to the use of phosphorus and sulphate-containing fertilizer (Pawellek *et al*., 2002).

8.5. Conclusions

The PMF model was successfully applied to the data set characterized by sub-basins with a different geological and urbanized impact. Three source factors were identified. Two factors explain the natural background influenced by the local geochemistry. A carbonates source was predominant in the upper part of the Danube, in concordance with the lithology of outcropping rocks, while an alumino-silicate component was mainly located in the last part of the river course

where loess deposits are abundant. Most of the measured heavy metals resulted bound to natural processes (alumino-silicate mineral and oxide phases) rather than to anthropogenic activities. The last resolved source was characterized by a potential anthropogenic impact mainly stemmed from wastewater discharge and mining activity. The factors spatial distribution map evidenced the role of tributaries: in the majority of tributaries locations the anthropogenic source shows a higher contribution. The application of the PMF model to the tributaries data set only, also identified a possible influence of fertilizer used in agriculture. Moreover the heavy metals content in tributaries sediments seems to be more connected to the anthropogenic activity than in the Danube sediments.

An interesting development could be achieved by performing further monitoring campaigns at the same sampling locations, in order to check possible changes in the river sediment sources. In particular, results here obtained can be used as a fingerprint of the Danube sediments status before a catastrophic event, i.e. the Hungary's red mud disaster happened in October 2010. Both two-way PMF or multi-way approaches, i.e. three-way PMF (Paatero, 2007a) and Multilinear Engine (ME-2; Paatero, 1999), could be applied to reveal possible hotspot contamination due to heavy metals accumulation following the red mud spill in Hungary.

Chapter 9

Nano-silver characterization

In the following experimental design, a protocol was developed and applied to study the quantification of silver in nano-form in wet samples, using inductively coupled plasma-atomic emission spectrometry (ICP/AES) technology and microwave assisted acid digestion. To this end, method validation procedure and budget uncertainty estimation were applied to test the accuracy of results. The total share of Ag could be in fact a key to develop reliable approaches, such as multivariate approaches, necessary for a large-scale assessment of nano-Ag environmental occurrence.

The choice of this methodology to analyze nano-Ag was based on the final goal to detect silver content in sewage sludge samples (Ch. 10). The first objective was the quantification of nano-silver in a representative reference material using ICP/AES and aqua regia microwave digestion, a procedure adopted for the determination of heavy metals in sewage sludge samples. The homogeneity of tested nanomaterial was then performed.

9.1. Nano-silver in the environment

Silver nanoparticles are most promising materials for a range of applications due to the property of silver to be an antibacterial and antimicrobial agent (Morones *et al.*, 2005; Kim *et al.*, 2007). Comparable types of uses are well known since a long time from medical applications and the field of biomedical devices. In biomedicine, vascular implants, such as coronary stents, catheters or orthopaedic devices have been designed using silver to better perform and function in its intended use and application (Laurin et al. 1987). Different nano-silver containing products were developed in other domains using its antibacterial activity, for example in recent applications as coating agent in textiles (Perelshtein *et al.*, 2008) or in wound dressing (Chen and Schluesener, 2008). According to the *Emerging Nanotechnologies database* (Woodrow Wilson International Center for Scholars, 2009), silver nanotechnology is present in more than 240 commercial products, ranging from medical applications, domestic appliances and cleaning products, antibacterial textiles, food storage and personal care products and also some kids toys. These new nano-silver-based products are nowadays part of everyday life, and hence in close contact with human beings and the environment. While nano-silver containing products provide

significant benefits due to their biocide effects, little is conclusively described about their environmental fate, toxicity and eco-toxicity, respectively (Handy *et al.*, 2008).

In the last years, some toxicity studies were carried out on aquatic species (Asharani *et al.*, 2008), human cells (Greulich *et al.*, 2009) and mammalian cells (Ahamed *et al.*, 2008, Arora *et al.*, 2009). Moreover, in a recent article (Kvitek *et al.*, 2008) the attention was also posed on the possible increase of silver nanoparticles ecotoxical effects by the interaction with surfactants/polymers.

A possible emerging problem is the risk due to the release of silver nanoparticles (NPs) directly into wastewater caused by the increasing use of household products containing nano-silver. In products the release of silver nanoparticles depends strongly on the method of fixation and embedding into the respective matrix. In contrast to nanosilver added during the initial fibre-spinning process, the simple functionalization of textiles by coating can in fact release silver during long time in their life cycle, like fabrics during regular washing (Benn and Westerhoff, 2008; Geranio et al., 2009), which is directly discharged into sanitary sewage system (Blaser *et al*., 2008; Benn and Westerhoff, 2008).

In **Fig. 46**, the silver flow released into wastewater is represented by Blaser *et al*. (2008). Wastewater from domestic sewer system enters a waste water treatment plant (WWTP) where the most nano-silver is removed and deposited in sewage sludge produced from waste treatment (Blaser *et al.*, 2008, Gottschalk *et al.*, 2009). Environmental contamination of silver can thus arise from the re-use of sludge, for example in agricultural soil, giving raise to soil and groundwater pollution (Blaser *et al.*, 2008). A modelling study concerning nanoparticles concentration in the environment, conducted by Mueller and Nowack (2008), reveals that the use of sludge as fertilised release about 1 $\mu g/kg^3$ nano-Ag per year, considering that 50% of agricultural land receives all sludge from WWTPs.

Considering the increasing use and developments of nano-silver household products, major silver NPs pathway becomes the sewer system. It is thus important to correctly quantify, as a first approach, the total silver content both in sludge and effluents from WWTPs; independently of its form (nano or not) it can in fact affect aquatic and terrestrial ecosystem. In order to reliably address the scientific questions of silver nanomaterials-induced effects, toxicity, ecotoxicity and fate, representative nanomaterials (NMs) are required, which are representative for industrial application and commercial use, for which a critical mass of study results are generated or known. These NMs will allow comparison of testing results, the development of conclusive

assessment of data, and pave the way for appropriate test method optimization, harmonisation and validation. They may serve as performance standards for testing.

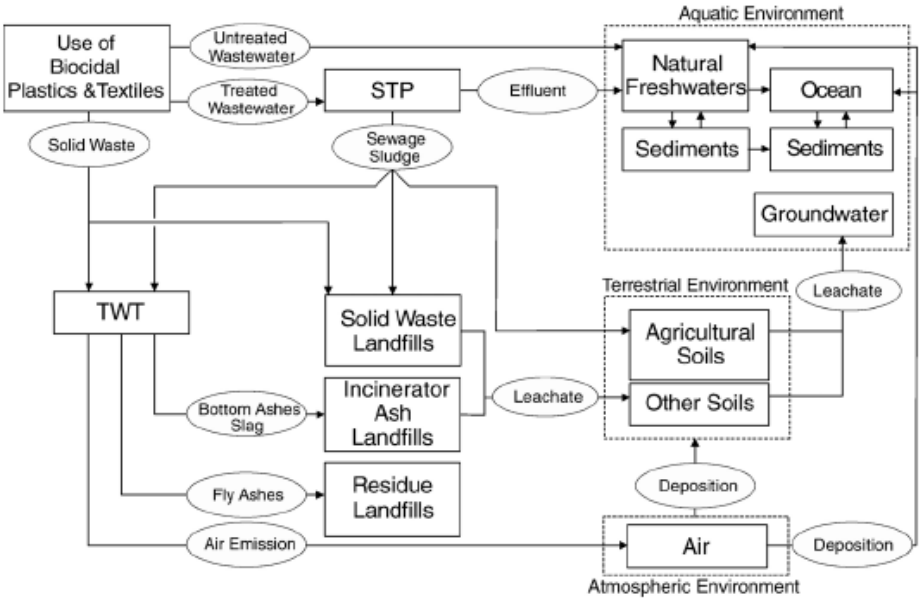


Fig. 46: Silver flows due to silver containing products (by Blaser et al., 2008).

In the following sections we address the silver content in a representative silver nano-material and the stability over a period of up to 12 months as well as homogeneity between vials.

9.2. NM-300 representative nanomaterial

The experiments were conducted using NM-300 nano-silver < 20 nm reference nanomaterial, used for measurement and testing for hazard identification, risk and exposure assessment studies. The further processed series of NM-300 is labelled with an additional "K" as NM-300K. It is a continued processed number of sub-samples from the same master batch of raw material. The material is a nano-Ag colloidal dispersion with a nominal Ag-content of 10 weight percent. The NM-300 appears orange-brown, yellow in dilution and consists of an aqueous dispersion of silver with stabilizing agents, 4% each of Polyoxyethylene Glycerol Trioleate and Polyoxyethylene (20) Sorbitan mono-Laurat (Tween 20). The ready material was distributed by the Fraunhofer Institute for Molecular Biology and Applied Ecology, Schmallenberg (Germany). Upon receipt at the Joint Research Centre, Ispra Site (Italy), samples were stored at 4ºC in the dark.

9.2.1. Handling procedure for weighing and sample introduction

A handling procedure has been established in cooperation with scientists at the different research institutions, which used the NM-300 and NM-300K, respectively. It takes into account that the material is a dispersion with a high amount of silver. The NM particles have the tendency to sediment slowly and should be homogenised within the vial before use by vigorously shaking the sample. Artefacts have been observed in a few cases consisting of larger aggregates or particles. In some cases, such aggregates were observed when the content of the NM vial was not discarded, but re-used. The NM vial contains an Argon atmosphere. If the vial is not kept upright or if remaining dispersion is drying at the edge of the vial, artefacts, such as larger aggregates may form. Dedicated sample and test item preparation protocols need to be used depending on the specific requirements of the measurement procedure or the test method.

The suggested handling protocol for NM-300 reads:

BE FAST, once the vial is open! If possible, work in a glove box under inert dry atmosphere.
The vial containing the NM material is filled with Argon. Keep the vial upright. Record the individual sample ID number as indicated on the NM label. If working outside glove box, please wear gloves.

1) Record laboratory conditions including relative humidity of the laboratory air for QA,
2) weigh a volumetric flask without cap,
3) Shake the vial before use: Make sure the vial is closed. Shake the vial vigorously for four minutes.
4) remove cap from the NM-300K material vial,
5) transfer an amount of dispersion into the volumetric flask using a pipette, determine and note down the weight of the volumetric flask with the transferred amount of NM-300K material,
6) close the NM-300K material vial,
7) calculate mass difference, which corresponds to the weight of transferred amount of NM-300K,
8) adjust to desired volume by adding Ultrapure (Type I) water quality as described in US-EPA, EP and WHO norms,
9) close the volumetric flask. Use this master stock dispersion for testing, accordingly.

General remarks:
A new pipette tip has to be used for each measurement.
Use Ultrapure (Type I) water quality as described in US-EPA, EP and WHO norms for dilution.
Store diluted samples in a refrigerator at 4 ºC in the dark, but keep time before use to a minimum.

9.3. Equipment

9.3.1. Inductively Coupled Plasma – Atomic Emission Spectroscopy

Silver analysis was carried out with the Optima 2100 DV ICP/AES device (Perkin Elmer) using the condition listed in **Tab. 16**. Silver concentration was determined in microwave assisted acid digestion solutions.

Tab. 16: operational conditions for ICP/AES.

Parameters	
Plasma condition	
Plasma flow (Argon)	15 L/min
Auxiliary flow (Argon)	0.2 L/min
Nebulizer flow (Argon)	0.8 L/min
Power	1300 W
View distance	15.0 mm
Plasma view	Axial
Peristaltic pump	
Sample flow rate	1.5 L/min
Autosampler	
Wash between samples for 30 s	

Silver was measured at the wavelength of 328.069 nm and peak area was used for the spectral peak processing. Silver ICP stock solution 1000 μg/mL, Ultra Scientific ICP-047, diluted in 2% nitric acid was used for ICP standard.

9.3.2.Microwave digestion

Microwave digestion of nano-Ag samples was performed by the Milestone 1600 device (Ethos). Microwave digestion conditions used in the study (**Tab. 17**) were previously optimized for sludge analysis using Certificate Reference Materials (CRMs).

Tab. 17: microwave program for nAg analysis. *Vent.* Stand for ventilation.

Time (min)	Power (W)
7	250
7	500
5	750
3	Vent.

For the digestion procedure, a 200 µL aliquot of diluted NM-300 solution was leached with 3 mL of 37% HCl Suprapure and 1 mL of 65% HNO_3 Suprapure. The type and amount of reagents used in the digestion reflect those of a previous study based on the optimization of mineralization procedure for sewage sludge matrix samples.

9.3.3.Density computation

To extract NM-300 sample aliquots, a gravimetric approach was preferred, since the nano-Ag material is very viscous. Indeed, during dilution procedure could happen that some droplets of solution remain inside the pipette tip. This procedure was also recommended as input in the general handling procedures of materials similar to NM-300.

In order to determine silver concentration, expressed in mg/kg, in the diluted solution it is necessary to compute the density of the NM-300 dispersion. To this end, a known amount of nano-Ag was collected and weighted for ten times. The average density was found to be **1.10 ± 0.03 kg/L**.

9.4. Method validation for quantitative silver determination by ICP/AES

Method validation for the analysis of total silver content in the NM-300 reference nanomaterial using ICP/AES technology and microwave assisted acid digestion, was conducted in compliance to ISO 17025 (ISO/IEC 17025: 1999).

9.4.1. Calibration study

A blank and five standard concentrations were analysed in three replicated for five different days in order to verify the linearity of the calibration curve. The five standard concentrations used for calibration were: 0.03 mg/L, 0.05 mg/L, 0.1 mg/L, 0.3mg/L and 0.5 mg/L.

Linear calibration curves (**Fig. 47**) and correlation coefficients (**Tab. 18**) were computed for each daily calibration.

Tab. 18: Regression coefficients of linear calibration curves.

	Day 1	*Day2*	*Day3*	*Day4*	*Day5*
Regression coefficient	0.9992	> 0.9999	> 0.9999	> 0.9999	> 0.9999

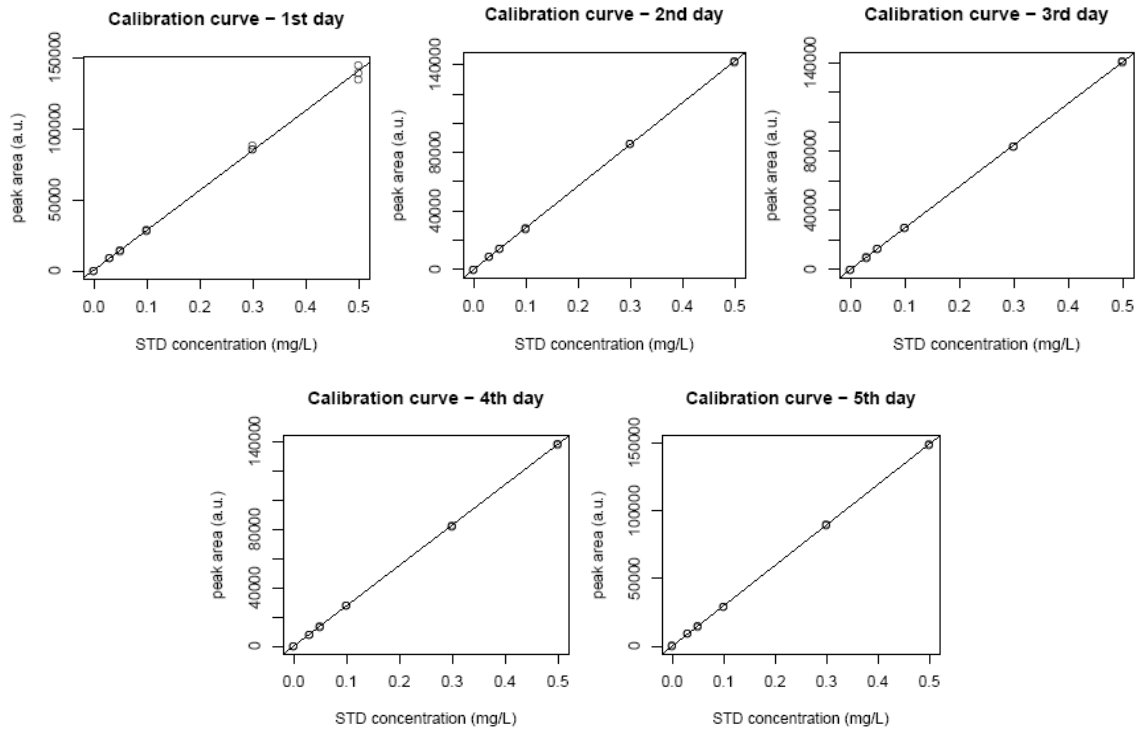


Fig. 47: ICP linear calibration curves for Ag.

Shape of calibration curves and regression coefficient values for each daily calibration prove the linearity and stability of the measurement system (Ag diluted standards and ICP instrument) in the 5-day range. Calibration curves were also obtained up to 9 days confirming the stability of the system. However, correlation coefficients are not the best parameter to prove the linearity (Loco *et al.*, 2002; González and Herrador, 2007). Means of residual plots (**Fig. 48**, the residual is the difference between the computed y-value and its prediction from the calibrating function) where thus used to confirm the linear trend. No trend is observed in **Fig. 48** for each daily calibration, confirming linearity.

Lack of fit test is another type of method used to test linearity. The way to do this is to determine if the modelling error is significantly different from the pure error, comparing the variance of the lack of fit against the pure error variance (González and Herrador, 2007.). The F-test showed that the linear calibration model was suitable for all the daily curves at the 99% confidence level; for the first two days the linear model adequately fit the calibration data at the 95% confidence level.

In order to test the homogeneity of variance, Bartlett's test and Fligner-Killeen's test were applied. The first method is more sensitive to non-normality of data, while the Flignet-Killeen's test is more robust in the case of departure from normality. Bartlett test reject the hypothesis of homogeneity of variance for all calibration curves while Fligner-Killeen's is significance for all calibration curves at the 95% confidence level.

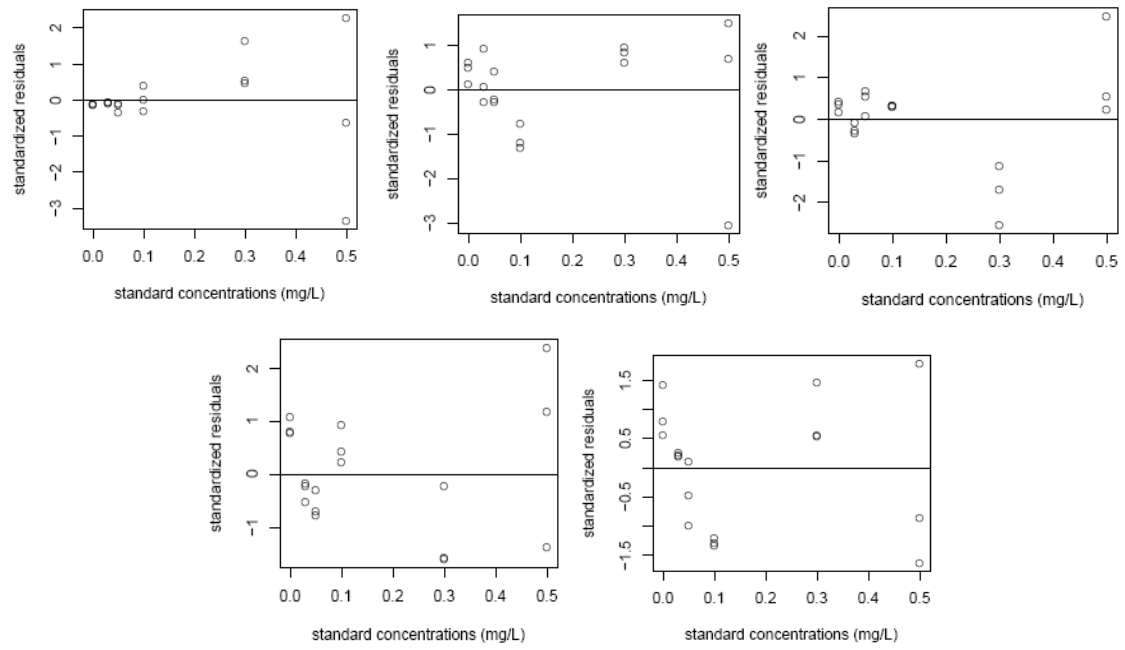


Fig. 48: residual plots for daily calibrations.

9.4.2.Working range

The working range is defined by the calibration curve (upper value) and the limit of quantification (LOQ). For higher concentration than those defined by calibration curve, a dilution is necessary.

9.4.3.LOD - LOQ

In order to estimate limit of detection (LOD) and limit of quantification (LOQ), a sample containing the selected analyte at very low concentration is analysed. Ten replicates were made in order to compute the standard deviation.

The following formulas were used to compute LoD and LoQ are:

$$\mathrm{LOD} = \Phi_{n,\alpha} \cdot \frac{s_L}{b}$$

$$\mathrm{LOQ} = k \cdot \Phi_{n,\alpha} \cdot \frac{s_L}{d} \qquad \textbf{Eq. 7}$$

where s_L is the standard deviation of the ten replicates. The $\Phi_{n,\alpha}$ factor takes into account the probability that certain response could be due to the standard deviation of the blank rather than the standard deviation of the analyte. The k factor corresponds to the reciprocal value of the desired accuracy. For 10 measurements and at a 95% confidence level ($\alpha = 0.05$) the $\Phi_{n\,\alpha}$ factor

is equal to 1.9. LOQ is computed using a *k* factor of 2, which give a 50% of accuracy. From this computation results:

LOD = 0.8 µg/L

LOQ= 1.6 µg/L

9.4.4.Trueness

Since certificate reference materials (CRMs) for nano-Ag material were not available, trueness was computed using the standard addition method. Two concentration levels of spiking were analysed in triplicate for five different days. To compute the recovery rates at each concentration level, three solutions were prepared: real sample, real sample with standard addition (*level 1*) and real sample with double standard addition (*level 2*). Daily Recovery rates are reported in **Tab. 19**. For the second level (*level 2*), during the 5th day, one replicate was rejected because of a suspected loss of sample during the rinse procedure after mineralization. The average recovery rate is 99%.

Tab. 19: daily recovery rates.

	day 1	*day 2*	*day 3*	*day 4*	*day 5*
level 1	100%	105%	103%	99%	100%
level 2	95%	100%	93%	98%	98%

9.4.5.Repeatability and intermediate precision

Repeatability and intermediate precision were computed analysing three samples at different concentration levels (low, medium and high) for 5 different days in three replicates. Results obtained from real and spiked solution in trueness evaluation were used.

Repeatability, intermediate precision (or within laboratory reproducibility) and day-to-day variation were evaluated using one-way analysis of variance (ANOVA). Results are presented in **Tab. 20** according to silver levels.

Tab. 20: repeatability and intermediate precision of ICP method for three silver concentration levels .

	low	*medium*	*high*
Repeatability	3 %	1 %	2 %
Intermediate precision	3 %	2 %	4 %
Day-to-day	2 %	1 %	1 %

9.4.6. Stability of the extracts

Sample extracted for trueness study were analysed for a week in order to check their stability. After one week, percentages of recovery do not vary significantly.

9.5. Estimation of the measurement uncertainty

The estimation of the measurement uncertainty was performed using the method expressed in the EURACHEM/CITAC Guide (Ellison *et al.*, 2000).

The aim of this uncertainty assessment was to provide the expanded uncertainty associated with the measurement of silver content in NM-300 material by ICP/AES techniques and microwave assisted digestion procedure. In order to analyse each source of error the cause-effect diagram was designed (**Fig. 49**). The combined uncertainty was computed using the propagation error law and the expanded uncertainty was obtained by multiplication of a coverage *k* factor, which takes into account the confidence limit (Ellison *et al.*, 2000).

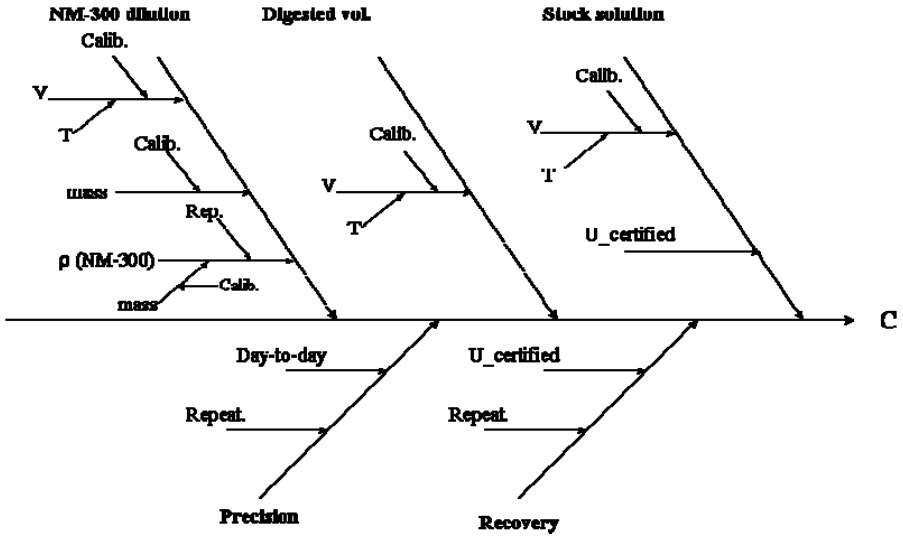


Fig. 49: cause-effect (or Hishikawa) diagram used for uncertainty assessment.

9.5.1.Combined uncertainty

The ICP/AES concentration of total silver content in each sample, was derived from the following equation:

$C = C_{ICP} \cdot d_1 \cdot d_2$ **Eq. 8**

C_{ICP} is the value, in mg/L, derived from ICP/AES analysis, and d_1 and d_2 are respectively the diluting factors before and after the mineralization process, expressed by:

$d_1 = V_1 \cdot \rho / m_{NM300}$

$d_2 = V_2 / \text{pipette}$

with m_{NM300} and ρ being the mass used for dilution and the density of NM-300 material, *pipette* the volume of diluted NM-300 solution used for mineralization process, $V_1$and V_2 the diluting volumes. Basing on the cause-effect diagram, the main factors that contribute to the overall uncertainty were found to be the method recovery, precision, concentration of diluted standard stock solution and NM-300, and the final volume of sample digest. Starting from the contribution of single uncertainties and using the error propagation low, the combined uncertainty expressed in terms of relative uncertainties u_i can be calculated using the following equation:

$$u_{rel}(C) = \sqrt{u_{rel}^2(\text{stock}) + u_{rel}^2(NM-300) + u_{rel}^2(\text{Vfinal}) + u_{rel}^2(\text{rec}) + u_{rel}^2(\text{precision})} \qquad \textbf{Eq. 9}$$

In the next sub-sections all these contributions are analyzed individually. The uncertainty due to pipetting operations was taken into account in the precision study because, during the NM-300 measurements, different fixed and adjustable-volume pipettes have been used.

Ag standard stock solution

The uncertainty associated with the silver stock solution used for calibration is a combination of the uncertainty associated with Ag content uncertainty, given in the standard solution certificate, and the uncertainty derived from the volumetric flask used for dilution.

The certificated standard uncertainty (given by the manufacturer) for Ag stock solution is 1000 ± 2 mg/L. Because this value is not correlated with a confidence level or distribution information, a rectangular distribution was assumed, dividing the uncertainty by $\sqrt{3}$

$u_{cert} = 2/\sqrt{3} = 1.15$ mg/L

The uncertainty of the volumetric flask (100 mL) used for the dilution of the stock solution was computed combining the uncertainties arising from temperature and calibration effects.

The tolerance of the volumetric flask, given by the manufacturer, is 0.1 mL at a temperature of 20 ºC. Since no confidence level is reported, a triangular distribution was assumed and the uncertainty associate with calibration effect was:

$U_{calib} = 0.1/\sqrt{6} = 0.04$ mL

In order to account for the temperature variability in the laboratory within ±3 ºC of the calibrating temperature (20ºC), a rectangular distribution was assumed and the uncertainty associate to this effect was computed with the following equation:

$$u_{temp} = \frac{T \cdot V \cdot Q}{\sqrt{3}} = 0.04 \text{ mL}$$

where T is the temperature variability (±3), V is the volume of the volumetric flask used and Q is the coefficient of volume expansion of the water (Q = 2.1 x 10^{-4} ºC^{-1}).

The combined uncertainty of the volumetric flask was then:

$$u_{volum} = \sqrt{u_{calib}^2 + u_{temp}^2} = 0.05 \text{ mL}$$

Tab. 21: combined uncertainty of Ag stock solution.

Description	*Value*	*SD*	*Uncertainty as RSD (%)*
Ag stock solution (u_{cert})	1000 mg/L	1.15 mg/L	0.12
Volumetric flask (u_{volum})	100 mL	0.05 mL	0.05
Combined uncertainty (u_{stock})	-	-	0.13

The combined uncertainty of the Ag stock solution was computed combining the uncertainties given in **Tab. 21**:

$$\frac{u_{stock}}{C_{stock}} = \sqrt{\left(\frac{u_{std}}{C_{stock}}\right)^2 + \left(\frac{u_{volum}}{V}\right)^2} = 0.13 \%$$

NM-300 diluted solution

The uncertainty associate to the NM-300 diluted solution is a combination of uncertainties arising from the volumetric flask, NM-300 mass and the density of the nano-material (ρ). The NM-300 silver content uncertainty was not known.

The uncertainty of the flask volume was already computed, being the flask used for NM-300 dilution of the same type of that used for Ag stock solution.

The contribution of NM-300 aliquots weight, used for dilution, is obtained from the uncertainty of balance linearity, reported in the calibration certificate. From balance linearity (± 0.03 mg), a rectangular distribution was assumed to compute the standard uncertainty; this contribution was considered twice (tare and gross weights). This gave, for the standard uncertainty of NM-300 mass u_m, the following value:

$$u_m = \sqrt{2 \cdot \left(\frac{0.03}{\sqrt{3}}\right)^2} = 0.02 \text{ mg}$$

The amount of NM-300 material used during the experiments was approximately the same. In order to calculate the relative standard deviation, the mean mass weight (55 mg) was considered. The uncertainty for NM-300 density is both due to the standard uncertainty in repeated measurements, u_{rep}, and in NM-300 mass. It was computed combining u_m and u_{rep}:

$$\frac{u_\rho}{\rho} = \sqrt{\left(\frac{u_m}{mass}\right)^2 + \left(\frac{u_{rep}}{\rho}\right)^2} = 2.88\ \%$$

Tab. 22: combined uncertainty of NM-300 diluted solution.

Description	*Value*	*SD*	*Uncertainty as RSD (%)*
Volumetric flask (u_{volum})	100 ml	0.05 ml	0.05
Mass (u_m)	55 mg	0.02 mg	0.04
Density (u_ρ)	-	-	2.88
Combined uncertainty (u_{NM300})	-	-	2.88

The combined uncertainty of NM-300 diluted solution was computed using the uncertainties given in **Tab. 22** as follow:

$$\frac{u_{NM-300}}{C_{NM-300}} = \sqrt{\left(\frac{u_{volum}}{V}\right)^2 + \left(\frac{u_{density}}{\rho}\right)^2} = 2.88\ \%$$

Final digested volume

This uncertainty is due to the 50 mL volumetric flask used to collect the sample after microwave digestion. The uncertainty associated with flask volume is, as already computed, a combination of calibration and temperature effects.

$$u_{Vfinal} = \sqrt{u_{calib}^2 + u_{temp}^2} = 0.04\ \text{mL}$$

Uncertainty expressed as relative standard deviation is reported in **Tab. 23**.

Tab. 23: uncertainty of volumetric flask for final digestion volume.

Description	*Value*	*SD*	*Uncertainty as RSD (%)*
Volumetric flask (u_{Vfinal})	50 mL	0.04 mL	0.09

Recovery

The overall bias of the analytical method is due to the recovery study determined in method validation. The uncertainty in recovery is derived from the standard deviation of the mean from

the trueness assessment study (u_{tr}) and from the uncertainty associated with the stock standard solution used for spiking:

$$u_{rec} = \sqrt{\frac{s_{tr}^2}{n_{tr}} + u_{std}^2}$$

where s_{tr} is the relative standard deviation derived from daily average recovery and $n_{tr} = 5$ is the number of days. Combined uncertainty contributions are given in **Tab. 24**.

The uncertainty associated with the Ag stock solution was previously estimated, assuming a rectangular distribution.

Tab. 24: combined uncertainty for recovery test.

Description	*Value*	*SD*	*Uncertainty as RSD (%)*
Trueness (u_{tr})	-	-	0.88
Ag standard (u_{cert})	1000 mg/L	1.15 mg/L	0.12
Combined uncertainty (u_{rec})	-	-	0.89

A t-test was computed in order to determine whether the mean recovery ($\overline{R}_m$) is significantly different from 1. The following parameter was estimated:

$$t = \frac{\left|1 - \overline{R}_m\right|}{u_{tr}}$$

The t value obtained was compared with the critical value, t_{crit}, with *n-1* degree of freedom at 95% confidence level (Ellison *et al*., 2000), where *n* is the number of results used to calculate the average recovery. Average recovery is significantly different from 1 if $t \geq t_{crit}$. It results that the average recovery is not significantly different from 1 ($t = 1.16 < t_{crit} = 2.04$) and no correction factor is to be applied.

Precision

Uncertainty associated with precision was derived from repeatability and intermediate precision computed in the validation study.

Uncertainty in repeatability was estimated as $s_{rep} / \sqrt{n_{rep}}$ where s_{rep} is the relative standard deviation due to repeatability experiment and n_{rep} the number of replicates.

The uncertainty due to intermediate precision was estimated as $s_{day} / \sqrt{d}$ with s_d being the relative day-to-day variation and *d* the number of days.

The precision uncertainty was derived combining these uncertainties:

$$u_{prec} = \sqrt{u_{rep}^2 + u_{day}^2}$$

Three concentration levels of uncertainty, low, medium and high, were computed (**Tab. 25**).

Tab. 25: uncertainty due to precision at different concentration levels.

Description	*RSD (%) Low*	*RSD (%) Medium*	*RSD (%) High*
Repeatability (u_{rep})	0.67	0.49	1.03
Intermediate precision (u_{day})	0.99	0.72	0.67
Combined uncertainty (u_{prec})	1.19	0.87	1.23

Total combined uncertainty

Starting from the contribution of single uncertainties, the combined uncertainty, expressed in terms of relative values u_i, can be calculated by the **Eq. 9**. All contributions to combined uncertainty are given in **Tab. 26**.

Tab. 26: relative standard deviation contribution for combined uncertainty.

Description	*Uncertainty as RSD (%)*
Stock solution (u_{stock})	0.13
NM-300 ($u_{NM\text{-}300}$)	2.88
Final Vol. (u_{Vfinal})	0.09
Recovery (u_{rec})	0.089
Precision (u_{prec})	0.87-1.23

From **Fig. 50**, it could be observed that the main contributions to the uncertainty estimation are NM-300 dilution (mainly due to NM-300 density uncertainty), method recovery and precision. The remaining two contributions, final volume and Ag stock solution, are not relevant.

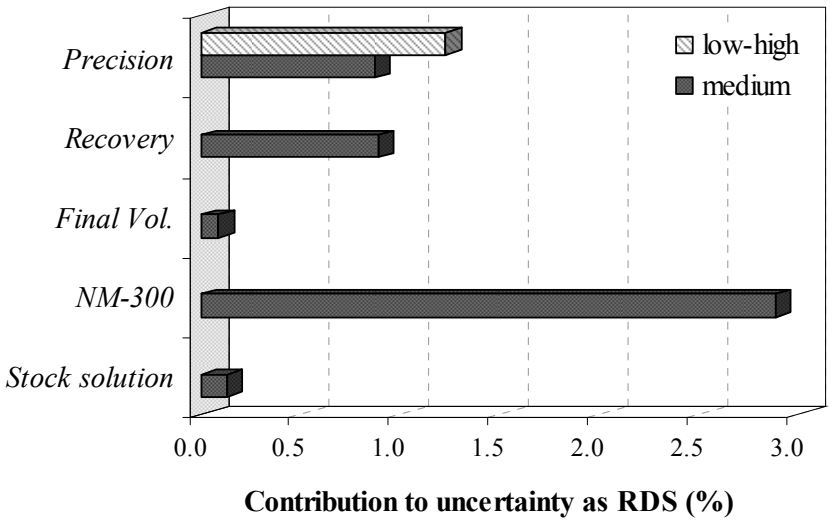


Fig. 50: major contributions to uncertainty.

The combined uncertainty was computed for two precision levels, low-high range and medium range:

$u_{combined}$ = 0.033 (medium) – 0.034 (low-high)

In percentage terms the combined uncertainty was found to be **3.3 - 3.4%**.

9.5.2. Expanded uncertainty

The expanded uncertainty was computed using a coverage *k* factor, starting from the combined uncertainty, in order to take into account a confidence level.

Taking into account a 95% confidence level, a *k*=2 factor was used. The expanded uncertainty, equal for both the precision levels, is given by:

$u_{expanded} = k \cdot u_{combined} = 0.07$

In percentage terms, the assessment of the uncertainty associated with the measurement of silver content in NM-300 material by ICP/AES and microwave assisted digestion, give an **expanded uncertainty of 7%**.

9.6. Homogeneity study

A homogeneity study was performed on the silver representative nano-material, in order to confirm the homogeneity of its silver content.

For this experiment, 20 units of NM-300 material were furnished. They were grouped in four categories according to their origin, as described in **Tab. 27**. Additional experiments addressed

within bottle homogeneity and the influence of the conditioning of the vial before sampling, including shaking the sample vigorously.

Tab. 27: NM-300 units used and conditioning groups.

	Sample ID	*Description*
Group 1	0030 - 0051 - 0021 - 0033 0075	*Samples from normal production process*
Group 2	0886 - 0897 - 1048 - 1170 1432 - 1468 – 1493	*Samples from normal production process, but different selection compared to group 1*
Group 3	5061 - 5062 - 5063 – 5064	*Samples from original homogenised (shaken) master-batch before sub-sampling*
Group 4	5065 - 5066 - 5067 – 5068	*Samples from original not homogenised (not shaken) master-batch before sub-sampling*

In order to obtain more suitable Ag concentrations for ICP/AES analysis, original NM-300 samples were diluted. Three independent sub-portions were collected from each NM-300 vial, using a 50 µL pipette, and diluted with Milli-Q water in a 100 mL volumetric flask. Starting from these diluted solutions, an aliquot of 200 µL was extracted for microwave assisted digestion. After digestion, the extract was diluted in 50 mL volumetric flask with Milli-Q water and analysed by ICP/AES. During the experiment, the NM-300 units had been re-homogenised prior to every sub-portion extraction, vigorously shaking the bottle for about four minutes.
The following scheme was applied in a four day time:

- *day 1*: first series of microwave assisted digestion on first sub-portion of NM-300 unit;
- *day 2*: second series of microwave assisted digestion on second sub-portion of NM-300 unit and first series analysis by ICP/AES;
- *day 3*: third series of microwave assisted digestion on third sub-portion of NM-300 unit and second series analysis by ICP/AES;
- *day 4*: third series analysis by ICP/AES.

A new ICP/AES calibration curve was daily performed. Digested samples were analysed in triplicate and mean values were used.
Silver concentrations from ICP analysis in tested sub-portions of NM-300 material are given in **Tab. 28**. It can be observed that mean concentrations of groups 1 and 2, both coming from the normal production process (but different batches) show different mean silver concentration. Moreover, silver-content in group 4, where vials come from the original not homogenises master-batch, shows the lowest mean value. A statistical t-test was applied for the comparison of

group 1 and 2 means and results show that they are significantly different at 95% confidence levels.

The obtained results were evaluated using one-way ANOVA. Between-bottle and within-bottle uncertainties (u_{bb} and u_w) were calculated according to the formulas used within ISO Guide 35 (1989).

Tab. 28: Silver concentration in NM-300 units derived from ICP/AES analysis for homogeneity study.

	Unit ID	***day 1*** *conc. (%)*	***day 2*** *conc. (%)*	***day 3*** *conc. (%)*	*Mean conc.(%)*	*Std. dev. (%)*
Group 1	0030	10.2	9.9	10.0	9.7	0.4
	0051	10.2	10.2	9.6		
	0021	9.5	9.9	9.1		
	0033	9.7	9.0	9.0		
	0075	9.4	9.5	9.6		
Group 2	0886	8.8	9.2	8.4	8.9	0.3
	0897	8.7	8.8	9.0		
	1048	8.8	8.6	8.7		
	1170	9.4	9.4	9.1		
	1432	9.1	8.7	8.8		
	1468	8.8	9.0	8.6		
	1493	9.2	9.2	9.3		
Group 3	5061	9.0	9.2	9.9	9.4	0.3
	5062	9.6	9.9	9.7		
	5063	9.1	9.5	9.1		
	5064	9.5	9.5	9.0		
Group 4	5065	8.6	8.7	8.6	7.9	0.5
	5066	8.2	8.1	8.0		
	5067	7.8	7.4	7.3		
	5068	7.3	7.5	7.4		

The between-unit (u_{bb}) and within-unit (u_{rep}) standard deviations, which represent relatively the homogeneity uncertainty and the repeatability, were calculated using the following equations:

$$u_{bb} = \sqrt{\frac{MS_{among} - MS_{within}}{n}} \qquad \textbf{Eq. 10}$$

$$u_{rep} = \sqrt{MS_{within}} \qquad \textbf{Eq. 11}$$

where MS_{within}, mean squares within the groups, and MS_{among}, mean squares among the groups, were derived from the ANOVA evaluation; *n* represent the number of test-portions for each unit. Results are shown in **Tab. 29**.

Tab. 29: Results from homogeneity test; $u_{bb}$and u_{rep} being the between unit and within unit standard deviation, respectively.

	u_{bb} (%)	u_{rep} (%)
Group 1	3.1	3.1
Group 2	2.3	2.2
Group 3	1.5	3.1
Group 4	7.2	1.7

From ANOVA results it is evident that only group 4 could be considered not homogeneous, with a between unit variation of 7.2%. For this group, samples were drawn from original master-batch containers, which were consciously not re-homogenized before sampling in order to simulate a process-related uncertainty and to further optimize the processes.

Starting from this finding, an additional experiment on 5 units of group 2 was carried out. The NM-300 units were settled for a week before a new ICP analysis. After a week two test portions were collected from each unit: one at the top and one at the bottom of the vial. Measurements were performed as previously described.

Although it was difficult to sample the two-level test portions at the same "depth" from each unit, results given in **Tab. 30** show a difference in silver content between the top and the bottom of the dispersion. Mean percentage difference between the two-level concentrations is 29%

Tab. 30: Results from homogeneity test. Top and bottom portions were taken after settled the units for a week, while mixed portions after shaking units for 4 minutes.

Unit ID	*upper (%)*	*lower (%)*	*Mixed (%)*
0886	8.9	11.3	9.2
1048	8.5	10.8	8.9
1432	8.1	12.5	9.2
1468	8.9	13.5	9.3
1493	8.4	12.6	9.7

These findings demonstrate that silver NPs tend to accumulate at the bottom of the vial over time, together with the stabilising agent. This is also confirmed by the different opacity top and bottom diluted portion, as illustrated in **Fig. 51**. Moreover, the t-test was used to compare mean values from the top and bottom aliquots, resulting in rejecting the hypothesis of equal means at 95% confidence level.

Fig. 51: top (right) and bottom (left) diluted test portions.

Finally, units used in this experiment were re-mixed for about four minutes and a test portion from the centre of the solution was taken in order to test if, after shaking, the NM-300 material returns to the original homogenisation level. Results are shown in the last column of **Tab. 30**. Mean difference in silver concentration between the re-mixed unit (**Tab. 30**) and the original, mixed, vials (**Tab. 28**) is 4%. Using the t-test, the hypothesis of equal means is accepted at 99% confidence level. This shows that the content of the NM-300 vial can be conditioned by vigorously shaking the vial for four minutes and that a homogenous within-bottle distribution will be reached by this treatment.

In order to test the homogeneity of NM-300K in a dedicated experiment, 8 vials were randomly selected. From each sample three independent sub-portions were collected on three different days. During the experiment, the NM-300K units had been re-homogenised prior to every sub-sampling, gently shaking the bottle for about four minutes. The above described working conditions were applied.

Silver content, derived from ICP analysis, in the three independent NM-300K sub-portions is reported in **Tab. 31**; also daily regression coefficients of calibration curves are shown. It is to note that Ag concentration values in each NM-300K unit were computed using the same density obtained for NM-300 material.

Tab. 31: Silver concentration in NM-300K units expressed in %.

	Day 1	***Day 2***	***Day 3***
Correl. coeffic.	> 0.9999	> 0.9999	> 0.9999
Sample ID	*conc. (%)*	*conc. (%)*	*conc. (%)*
0078	8.4	9.0	8.8
0079	8.9	9.3	9.1
0082	8.8	9.3	9.4
0085	8.6	8.8	9.1
0095	9.0	9.3	10.5
0103	8.8	10.3	9.6
0110	9.7	9.8	10.3
0117	8.5	8.2	8.8

The obtained results were evaluated using one-way ANOVA. Between-bottle and within-bottle uncertainties (u_{bb} and u_w) were computed using **Eq. 10** and **Eq. 11**, respectively.

Results for between unit and within unit standard deviation are:

$\boldsymbol{u_{bb} = 4.6\ \%}$ $\boldsymbol{u_w = 4.9\ \%}$

indicating the homogeneity of silver content in NM-300k units.

In conclusion, for NM-300 the mass-related content determined by ICP/AES was 9.7 % with 0.4 % standard deviation. For NM-300K the mass-related content was 9.2 % with 0.6 % standard deviation. From t-test, the silver content in both batches was statistically not different.

9.7. Conclusions

ICP/AES and microwave assisted acid digestion, with optimized conditions, could be used to detect total silver-content, here in nano-form, in NM-300 and NM-300K material with an expanded uncertainty of 7%. This is a first step experiment for the determination of total silver, including the nano fraction, in complex matrices, like sewage sludge samples.

Being that the 'nano' characteristic of silver is no longer maintained when this fraction reaches the environment (Geranio *et al*., 2009; Kim *et al*., 2010; U.S. EPA, 2010), further approaches than its quantification by analytical chemistry instruments, have to be studied. In the following chapter a first effort to characterize nano-silver in environmental samples by multivariate modelling (i.e. PMF) was carried out.

Chapter 10

Application 4 - FATE-SEES project

The FATE-SEES project is a European monitoring campaign carried out by JRC-IES aiming at the determination of principal inorganic and organic pollutants that can affect effluent waters and sewage sludge produced by European WWTPs. These WWTPs end-products are of great interest because of their disposal reuse and reclamation. Level of pollutants must be monitored in order to prevent environmental pollution. In particular, sewer system became a major pathway for engineered silver NPs because of the increasing commercial of house-hold and personal care product containing nanosilver technology.

Within this project, we also optimized a method to evaluate total-silver content in both effluents and sewage sludge samples. The idea was to use a multivariate statistical approach in combination with traditional analytical chemistry to determine a nanosilver-related source in these compartments. However, being that silver concentrations is under the method limit of detection in all effluents samples, PMF was carried out only on sewage sludge samples.

10.1. Effluents campaign

A total of 91 effluent samples were collected across 18 European countries. A summary of number of samples collected in each country is given in **Tab. 32**. Sampling locations are illustrated in **Fig. 52**. Missing coordinates were found in Belgium (2) and France (3).

Upon sample receive, an aliquot was extracted and filtered through a 0.45 μm pore diameter membrane filter in order to determine dissolved elements. The filtrate was then acidified at pH<2 using nitric acid.

Tab. 32: number of effluents samples collected in each country.

Country	*N. samples*	*Country*	*N. samples*
Austria	6	Ireland	2
Belgium	18	Italy	2
Cyprus	2	Lithuania	3
Czech Republic	7	Portugal	2
Finland	6	Slovenia	1
France	5	Spain	3
Germany	3	Sweden	11
Greece	2	Switzerland	5
Hungary	2	The Netherlands	11

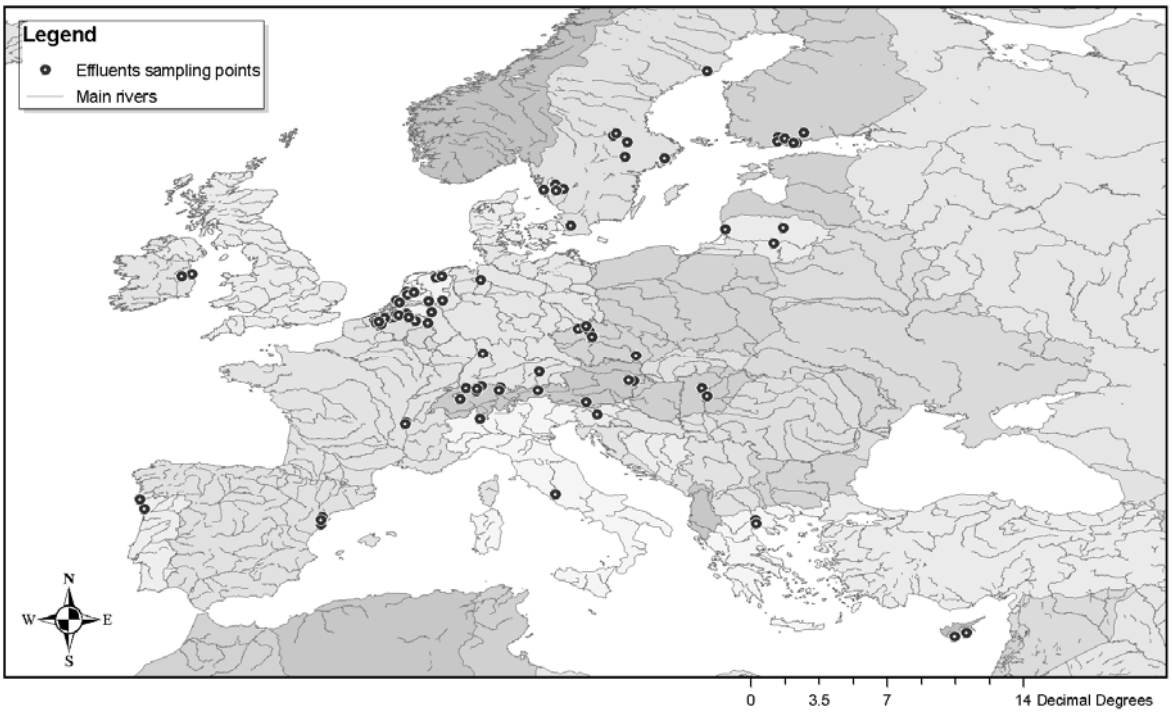


Fig. 52: map of collected WWTP effluent samples.

Major and minor elements, and heavy metals were determined by ICP/AES (Optima 2100 DV, Perkin Elmer) on filtrate aliquots: Ag, Al, As, Ba, Be, Cd, Co, Cr, Cu, Mg, Mn, Mo, Ni, Pb, Sb, Se, Tl, and Zn. Total mercury was determined by CV/AAS technique (AMA 254, FKV).

Single elements stock standard solutions were opportunely diluted to obtain standards for calibration both in the low and in the high range of concentration. The ICP operating conditions were the same used for nano-Ag detection (**Tab. 16**).

Majority of determined elements, including heavy metals, were not detected (BDL). A summary of detected concentrations and frequencies of detection is reported in **Tab. 33**.

Together with inorganic determination, organic compound were determined at JRC-IES laboratories and external European laboratories.

Due to the poor number of samples with detectable elements concentration no statistical analysis was carried out on inorganic data. Results reveal the limits of applicability of ICP/AES technique, using the condition listed in **Tab. 16** for measurement of low chemical concentrations (µg/L order).

Tab. 33: frequency of detection, minimum and maximum concentrations for elements detected in effluents samples.

	Frequency	Min. (mg/L)	Max. (mg/L)		Frequency	Min. (mg/L)	Max. (mg/L)
Hg	0%	-	-	***Cu***	2%	0.026	0.030
Ag	0%	-	-	***Mg***	100%	0.106	144
Al	9%	0.046	0.58	***Mn***	54%	0.005	0.49
As	0%	-	-	***Mo***	8%	0.011	0.50
Ba	25%	0.006	0.051	***Ni***	3%	0.051	0.42
Be	0%	-	-	***Pb***	0%	-	-
Cd	0%	-	-	***Sb***	1%	0.90	0.90
Co	1%	0.065	0.065	***Se***	0%	-	-
Cr	0%	-	-	***Zn***	84%	0.007	0.24

10.2. Sewage sludge campaign

A total of 61 samples were collected in 15 European countries. Some sewage sludge samples were collected at the same WWTP facilities of effluent campaign. Number of samples collected in each country is summarised in **Tab. 34**. Map of collected samples is illustrated in **Fig. 53**; some missing coordinates were found in Belgium (4) and Switzerland (1).

Tab. 34: number of sludge samples collected in each country.

Country	***N. samples***	***Country***	***N. samples***
Austria	2	Lithuania	3
Belgium	9	Portugal	2
Czech Republic	2	Romania	1
Finland	6	Slovenia	1
Germany	6	Sweden	8
Greece	3	Switzerland	9
Hungary	1	The Netherlands	6
Ireland	2		

Samples were analysed using ICP/AES technique (Optima DV 2100, Perkin Elmer) and microwave assisted acid digestion. The following elements were determined: Ag, Al, As, Ba, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Ni, P, Pb, Sb, Se, Ti, V and Zn. Mercury analysis was performed by CV-AAS technique, using an AMA 254 device (FKV).

An analogous campaign was carried out by U.S. Environmental Protection Agency (U.S. EPA) between 2006 and 2007. Within the Target National Sewage Sludge Survey (TNSSS), 84 treated sewage sludge samples were collected in 74 Publicly Owned Treatment Works (POTWs) located in the Unites States (U.S. EPA, 2009). All samples were analysed for 145 pollutants, including

both organic and inorganic compounds. In particular 28 metals, including mercury, were detected by ICP/AES, ICP/MS and CVAA techniques.

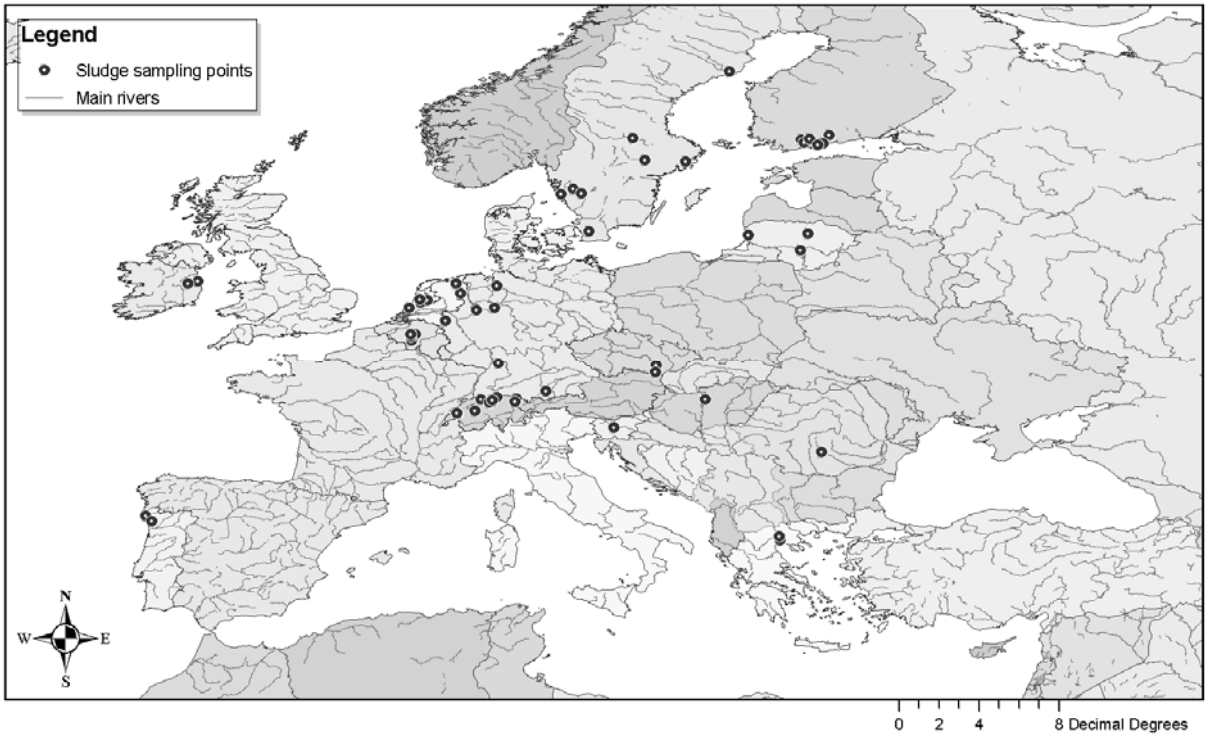


Fig. 53: map of collected WWTP samples.

10.2.1. Method

Prior to mercury and elements determination, all samples were freeze-dryed using a Gamma 1-16 LSC device (Martin Christ). Freeze-dried samples were then gently grounded in mortar with pestle to obtain more homogeneous powders.

CV/AAS analysis

Mercury was determined on freeze-dried samples using CV-AAS technique. The operating conditions are listed in **Tab. 35**. Single mercury stock standard solutions were opportunely diluted to obtain standards for calibration both in low and high concentration ranges. From three to five test portions were analysed to check the sample homogeneity.

Tab. 35: CV/AAS operating conditions for sludge samples analysis.

Parameter	*Time*
Drying time	60s
Decomposition time	200s
Cuvette clear time	45s
Delay	0s
Cell to use for analysis	Low / High cell
Metric to use for calculation	Peak area

ICP/AES analysis

Major and minor elements, and heavy metals were determined by ICP/AES after microwave assisted acid digestion treatment. Due to the high amount of samples to be measured the old microwave device, used in n-Ag experiment, was replaced. The new device, a Multiwave 3000 microwave (Anton Paar) was optimized for sewage sludge analysis. The microwave autoclave can simultaneously digest up to 48 samples in the reaction chamber under identical experimental conditions. One to three test portions were digested, depending on sample homogeneity. Mercury content was chosen as homogeneity parameter control: if the three-to-five-replicates relative standard deviation for mercury analysis was lower than 10%, one test portion was used in ICP/AES determination; three otherwise.

About 0.1 g of sludge sample was mixed with 1.5 ml of HNO3 and 4.5 ml of HCl, in the high-pressure, closed, Teflon decomposition vessel. The optimised program for sludge samples is listed in **Tab 36**.

Tab. 36: operating condition for microwave assisted acid digestion.

	Power (W)	*Ramp (mm:ss)*	*Hold (mm:ss)*
1.	1225	05:00	35:00
2.	ventilation	-	05:00

maximum IR temperature = 140ºC
maximum vessel pressure = 20 bar

After digestion procedure, each extract was filtered in a 50 ml glass flask using a clean glass funnel and 0.45 μm pore size filters. Vessel and the vessel cup were subsequently rinsed three times with Milli-Q water and the rinse water was filtered in the same flask. At the end, the flask was completed to volume and samples were stored at 4 ºC until analyses.

Single elements stock standard solutions were opportunely diluted to obtain standards for calibration both in the low and in the high range of concentration. The ICP operating conditions were the same used for nano-Ag detection (**Tab. 16**).

10.2.2. Method Validation

ICP/AES and CV/AAS methods used for the analysis of major and minor elements, heavy metals and mercury in sewage sludge samples were validated according to the ISO 17025 requirement. The same statistical tests used in nano-Ag validation were applied; only numerical results were here summarized.

For ICP/AES measurement, low and high calibration ranges were defined by 0.02-0.5 mg/l and 0.5-5 mg/l, respectively. Correlation coefficients were higher than 0.999 for the five-day linearity check, in both calibration ranges. The linear model adequately fit the calibration data at the 99% confidence level (lack-of-fit test). The homogeneity of variance, tested with Fligner-Killeen's test, was assumed with 95% confidence.

For mercury determination, low and high calibration curves were set to 0.05-0.5 mg/l and 0.5-5 mg/l, respectively. Correlation coefficients were higher than 0.995 and 0.996 in low and high range, respectively. By lack-of-fit test, the linear model was satisfactory only for some calibration curves, while the quadratic model adequately fit all the calibration data at the 95% confidence level. The homogeneity of variance, tested with Fligner-Killeen's test, was assumed with 95% confidence.

The working range for all measured elements was defined by the high calibration curve (upper value) and the limit of quantification (LOQ). For higher concentration than those defined in calibration, the measured solution has to be diluted and re-analysed.

LOD and LOQ were determined using the formula expressed in **Eq. 7**. Results, listed in **Tab. 37**, were expressed in mg/kg dry weight.

Tab. 37: LOD and LOQ determined by ICP/AES and CV/AAS (mercury only). Results are expressed in mg/kg dry weight.

	Hg	*Ag*	*Al*	*As*	*Ba*	*Cd*	*Co*	*Cr*	*Cu*	*Fe*	*Mg*
LoD	0.004	0.06	1.53	2.63	0.02	0.09	0.18	0.16	0.19	6.66	3.58
LoQ	0.008	0.12	3.06	5.25	0.04	0.18	0.35	0.32	0.38	13.32	7.15

	Mn	*Mo*	*Ni*	*Pb*	*Sb*	*Se*	*Ti*	*V*	*Zn*	*P*	*K*
LoD	0.02	0.36	0.14	1.26	1.66	1.78	0.03	0.81	2.12	3.03	4.83
LoQ	0.03	0.72	0.27	2.52	3.32	3.56	0.05	1.62	4.23	6.06	9.66

Trueness was determined using *CNS311-04-050* and *LCG-6181* certified reference materials and spiking solutions when element concentrations were not available in CRMs.

Average recovery, obtained in the 5 days calibration, for low and high ranges are listed in **Tab. 38**.

Tab. 38: Element recoveries, expressed in %, for low and high calibration ranges.

	Hg	*Ag*	*Al*	*As*	*Ba*	*Cd*	*Co*	*Cr*	*Cu*	*Fe*	*Mg*
Low	103%	101%	-	83%	89%	95%	98%	96%	93%	-	-
High	117%	92%	103%	90%	95%	88%	89%	98%	99%	98%	96%

	Mn	*Mo*	*Ni*	*Pb*	*Sb*	*Se*	*Ti*	*V*	*Zn*	*P*	*K*
Low	87%	87%	97%	94%	101%	83%	90%	99%	-	-	-
High	92%	92%	96%	97%	91%	92%	92%	93%	89%	122%	102%

For Al, Fe, Mg, Zn, P and K elements only high recoveries were determined because their concentration in sludge sample is usually high.

Repeatability, intermediate precision and day-to-day variation were evaluated, for both low and high concentration level, using one-way ANOVA. Values range between 1% and 11% depending on the selected element. Single results are provided in **App. A.1** and **A.2**.

10.2.3. Uncertainty

The expanded uncertainty for mercury, major and minor elements and heavy metals detection was estimated according to the guide EURACHEM/CITAC Guide CG4 (2000). For a detailed description of the procedure followed refer to Ch. 9. Here, only summary data were reported. In order to define each source of error, the cause-effect diagram was represented (**Fig. 54**).

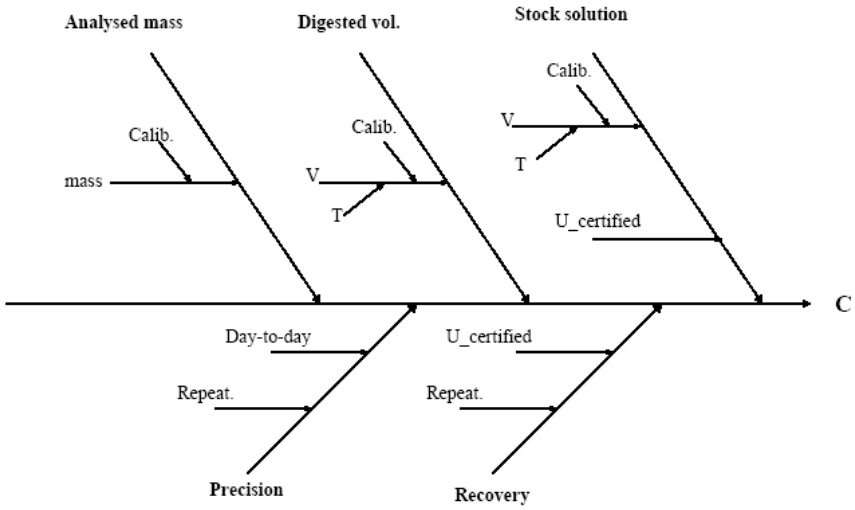


Fig. 54: Ishikawa diagram used for heavy metal content assessed by ICP/AES and microwave assisted acid digestion.

Basing on the cause-effect diagram, the main factors that contribute to the overall uncertainty were found to be the method recovery, precision, concentration of diluted standards stock solutions and the final volume of sample digest (except for mercury, which was determined on freeze-fried samples). Starting from the contribution of the single uncertainties, expressed in terms of relative uncertainties u_i, the combined uncertainty could be computed using the error propagation low.

The uncertainty due to recovery is derived from the standard deviation of the mean of trueness assessment study. Usually, also the uncertainty associated with the nominal value of CRMs is taken into account. However, in this case both CRMs and spiking solution were used for recovery study and, being the uncertainty associated with spike lower than the nominal uncertainty of CRMs, elements could show very different uncertainty ranges. Moreover, large uncertainties in nominally CRMs values could have a high impact on the overall uncertainty, making comparability very poor (Barwick and Ellison, 1999). In order to have more comparable data it was chosen not to use this term in the uncertainty formula. In **Tab. 39** all this single uncertainty contributions are summarized.

Tab. 39: uncertainty contributions expressed as relative standard deviation (RDS).

	Uncertainty as RDS (%)	
Description	*ICP/AES elements*	*Mercury*
Elements standard stock solutions	0.13	0.08
Mass used for microwave digestion / for mercury determination	0.02	0.18
Final digested volume	0.09	-
Recovery (element depend)	1 – 4	2.2 (low) – 1.8 (high)
Precision (element depend)	1 – 4	2.2 (low) – 3.3 (high)

It can be observed, from **Tab. 39**, that recovery and precision are the major contributions to the uncertainty budget.

The expanded uncertainty was computed multiplying the combined uncertainty by a coverage factor of 2. Values computed for sludge samples in low and high ranges of concentration are listed in **Tab. 40**.

Tab. 40: Expanded uncertainty (%) for ICP/AES measured elements and mercury.

	Hg	***Ag***	***Al***	***As***	***Ba***	***Cd***	***Co***	***Cr***	***Cu***	***Fe***	***Mg***
Low	6.1	4.5	-	6.3	6.1	5.6	7.1	6.0	3.0	-	-
High	7.5	6.2	7.2	4.0	6.7	5.5	5.0	1.3	5.8	5.2	7.9

	Mn	***Mo***	***Ni***	***Pb***	***Sb***	***Se***	***Ti***	***V***	***Zn***	***P***	***K***
Low	3.9	3.9	6.0	6.9	5.5	3.3	8.3	5.3	-	-	-
High	6.9	3.5	2.6	2.3	10	9.3	11	4.0	4.0	8.6	7.7

10.2.4. Statistics

A descriptive statistic of measured variables is given in **Tab. 41**. A high percentage of below-detection-limit data was found for the elements As, Sb and Se and they were excluded from further multivariate analysis. In the computation of statistic parameters, uncensored data were used. However, when negative data occurred, they were replaced with the DL/2 estimate. Missing value were found only for K and P and where manage in the data set by they average value.

Tab. 41: Descriptive statistic for measured elements. Percentages of below-detection-limit and missing data are also shown.

	Min	***Max***	***Mean***	***Median***	***STD***	***CV (%)***	***Skewness***	***BDL (%)***	***MV (%)***
Hg (mg/kg)	0.10	1.13	0.45	0.41	0.23	52	0.88	-	-
Al (%)	0.07	5.97	1.67	1.30	1.19	71	1.80	-	-
Ag (mg/kg)	< 0.06	14.7	3.30	2.37	3.01	91	1.67	5	-
As (mg/kg)	< 2.63	56.1	5.61	2.91	8.26	147	4.26	66	-
Ba (mg/kg)	41.5	580	225	197	102	45	1.06	-	-
Cd (mg/kg)	< 0.09	5.11	0.93	0.86	0.70	75	3.89	7	-
Co (mg/kg)	1.54	16.7	6.26	5.61	3.33	53	1.12	-	-
Cr (mg/kg)	10.8	1542	79.8	37.9	215	269	6.06	-	-
Cu (mg/kg)	27.3	578	257	240	118	46	0.48	-	-
Fe (%)	0.22	14.9	3.82	2.45	3.61	94	1.46	-	-
K (%)	0.10	2.57	0.43	0.36	0.36	85	3.82	-	11
Mg (%)	0.01	2.24	0.44	0.37	0.33	76	3.00	-	-
Mn (mg/kg)	75.2	960	329	281	193	59	1.31	-	-
Mo (mg/kg)	1.73	12.5	4.95	4.97	1.90	38	1.00	-	-
Ni (mg/kg)	8.64	310	29.0	20.1	40.2	139	6.10	-	-
P (%)	1.00	5.64	3.14	3.09	1.08	34	0.31	-	11
Pb (mg/kg)	3.96	430	47.6	30.4	59.3	125	4.82	-	-
Sb (mg/kg)	< 1.66	53.6	5.99	3.89	8.23	137	4.50	34	-
Se (mg/kg)	< 1.78	7.42	1.01	0.89	0.88	87	6.72	98	-
Ti (mg/kg)	65.2	1071	440	350	255	58	0.66	-	-
V (mg/kg)	2.35	135	25.0	21.5	20.3	81	3.05	-	-
Zn (%)	0.02	0.12	0.07	0.07	0.02	35	0.06	-	-

A high positive skewness coefficient was found for the majority of elements, indicating the presence of possible outliers. Outliers could be due to local hotspot, being the population of samples very different. A high coefficient of variation is also expected, because samples came from different WWTPs situated in several European countries. In particular, from boxplot (**Fig. 55**) it can be shown that elements with greatest variation are As, Sb, Cr and Fe.

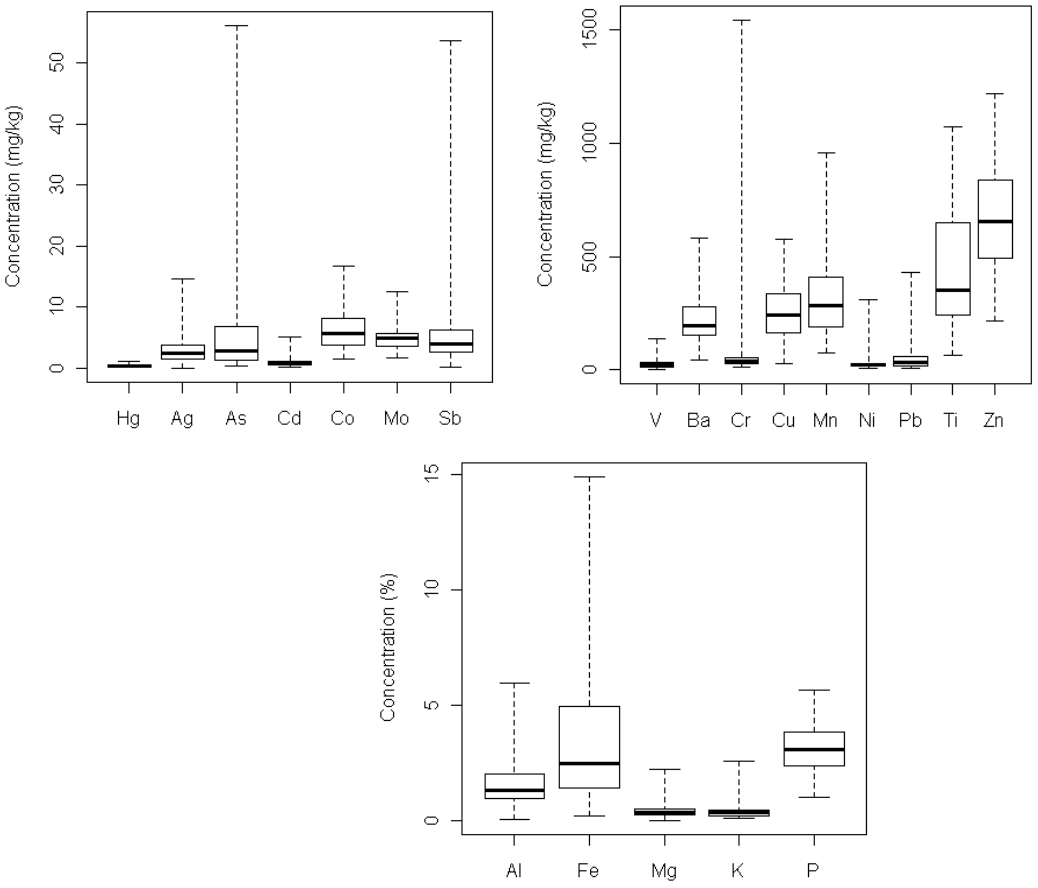


Fig. 55: Boxplot of measured elements. Se is not shown.

As mentioned before, in the U.S. EPA-TNSSS campaign, 28 metals were analyzed in sewage sludge samples. Since in the EPA report average values were not provided for all the listed elements, it was useful to compare minimum and maximum valued from United States POTWs and European WWTPs. The main differences between the two projects reside in the type of WWTPs considered. In the U.S. EPA survey, only municipal WWTP were considered, while in the FATE-SEES campaign both industrial and municipal facilities were examined. Moreover statistic was made on a different number of samples: 74 in United States and 61 in Europe. Comparison graph for common elements measured in the EPA and FATE-SEES campaigns are shown in **Fig. 56**.

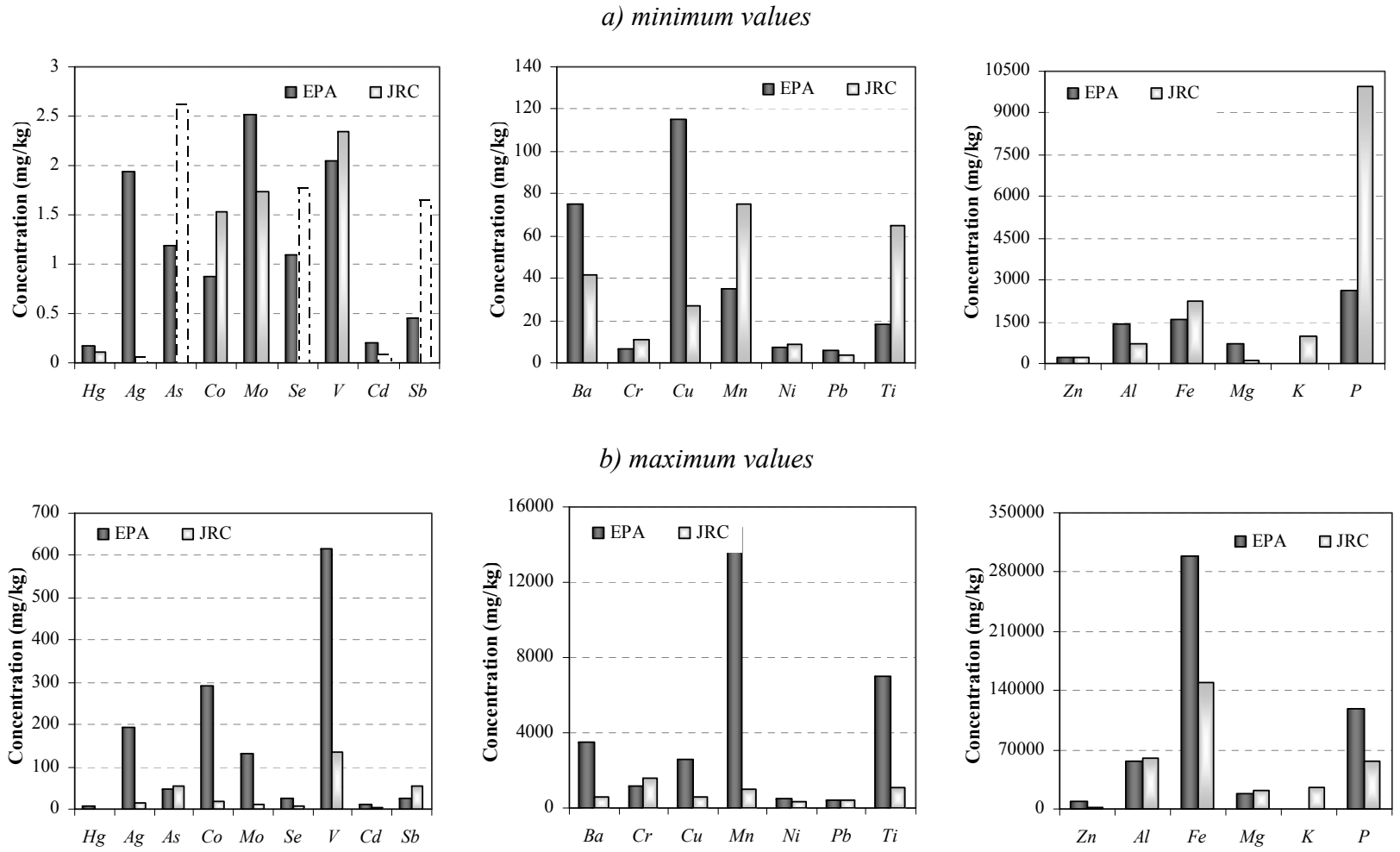


Fig. 56: Comparison between a) minimum and b) maximum values from FATE-SEES campaign and EPA TNSSS project. Dotted white boxes for FATE-SEES data represent the detection limit value.

In the European Union, the regulation of the use of sewage sludge in agriculture is defined in the *Directive 86/278/EEC*. Limit values for heavy metal concentrations were fixed for Cd, Co, Ni, Pb, Zn and Hg. These concentration limits were compared with results obtained in sewage sludge analysis (**Tab. 42**). Only in one WWTP the Ni concentration was found in the range of regulatory limits. For all other metals, the maximum measured concentrations were well below the regulation limit values.

Tab. 42: Limit values for heavy metal concentration in sludge for use in agriculture (Directive 86/278/EEC) and mean and maximum concentrations found in sewage samples. Values are expressed in mg/kg of dry matter.

Analyte	*Limit values*	*Mean conc. in sewage samples*	*Max conc. in sewage samples*
Cadmium	20 to 40	0.93	5.11
Copper	1000 to 1750	257	578
Nickel	300 to 400	29	310
Lead	750 to 1200	48	430
Zinc	2500 to 4000	663	1218
Mercury	16 to 25	0.45	1.13

10.2.5. PMF analysis

The As, Se and Sb were omitted from the analysis because of the high percentage of below-detection-limit data (**Tab. 41**). For silver and cadmium, which show <10% of BDL data, the uncensored values for BDL were used in the analysis. Potassium and phosphorus show some missing values, which were substituted by their average concentration.

The error estimate matrix was built using the error model EM= -14 with the following parameter: T is the matrix of LOD and V the matrix of uncertainties, both computed during method validation. For BDL data the uncertainty was doubled, while for MV the uncertainty value was multiplied by 4.

Initially, PMF2 was run varying the number of factors from 2 to 10. Q values, MaxRotMat, IM and IS parameters derived from the analysis are reported in **Fig. 57** and **Fig. 58**.

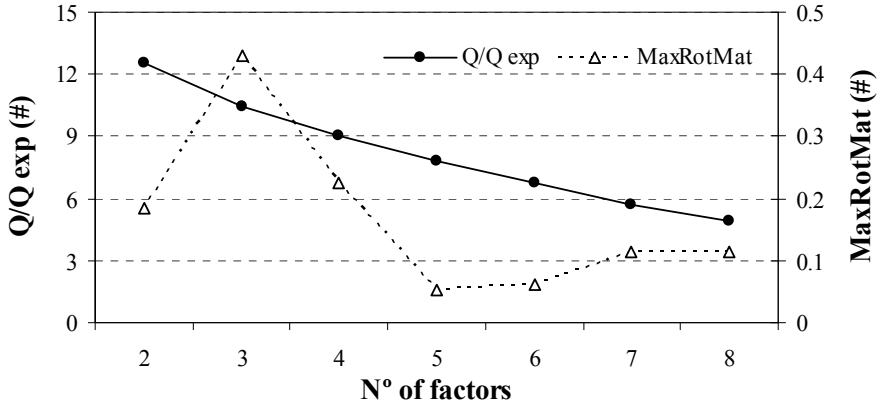


Fig. 57: Q vs. Q expected and MaxRotMat parameters for each number of factors examined.

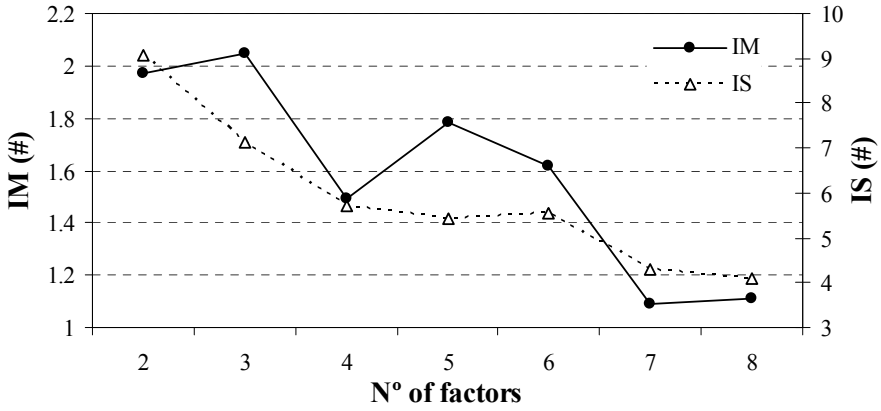


Fig. 58: IM and IS parameters values for each examined number of factors.

The Q value is decreasing along all the factors, while the MaxRotMat parameter has maximum values at 3 factors extracted. IM and IS have a first decreasing step from 4 to 7 factors, which is more evident for IS parameters. Solutions with more than 6 factors were excluded from further analysis taking also in consideration NEVF values for the measured variables. In fact, more is the number of factors resolved and more is the number of variables which are explained by a unique factor. This could describe the data set variability, i.e. for variables which are marker from a certain source, but could also arise from a too high number of factors selected.

Rotations were evaluated for solution with 4, 5 and 6 resolved factors, with the FPEAK parameters ranging between -1 and +1. For all the explored number of factors, the rotated Q do not differ significantly from the central Q (less than 2%). However, for 5 and 6-factor solutions, the IM and IS parameters show for some rotations a strong variation, up to 30% from the central value (**Fig. 59**).

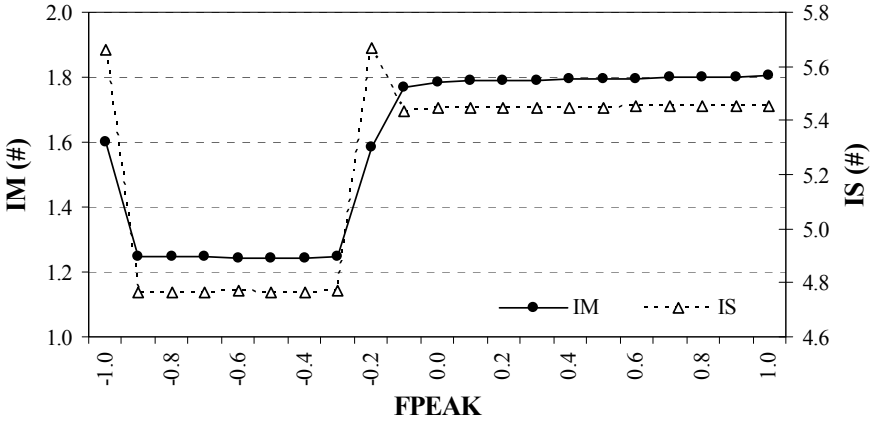


Fig. 59: IM and IS plot for the 5-factors solution.

These variations are consistent with a sharp change in the factors explanation within the same solution. That is, in 5 and 6-factors solutions, EVF assume different values.

With 4-factors resolved instead, the solution is more stable with IM and IS parameters, and also EVF values being more comparables. No significant changes resulted in varying *FPEAK* parameters and G-plots evaluation gave satisfactory results for all the rotations. The 4-factor solution was chosen, because it reflected more stable data. With more than 4 factors extracted no beneficial effects were observed, being probably the additional factors caused by the isolation of single variables in unique factors; this could be due to the strong data variability within the data sed. Indeed, we have to keep in mind that sludge samples were collected from WWTPs in different European countries. Factor resolution must be consisted with sources or processes common to all the selected facilities. It could thus happen that trying to force the model to explain more factors, hotspots were isolated in unique factors.

The 4-factor central solution was chosen; EVF values characterizing the source explanation are reported in **Fig. 60**.

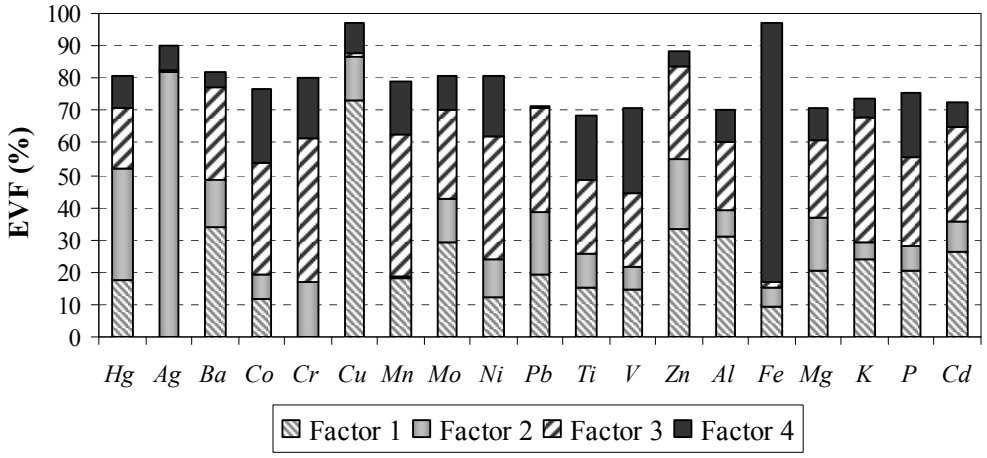


Fig. 60: Explained Variation of F for the 4-factor solution with FPEAK=0.

Factor 1

Factor 1 is mainly characterized by Cu variation. Copper was found in many studies to be connected with the corrosion of domestic water pipe lines (Fjällborg and Dave, 2003; Fabbricino *et al*., 2005; Houhou *et al*., 2009). This element is in fact a well know plumbing material. Copper source here identified could be associated with Cu dissolution from the inner surface of a pipe by tap water.

Factor 2

This factor is mainly explained by Ag variation and, to a lower extent by Hg. The association between Ag and Hg may be due to their common behaviour with sulphur: both the elements tend in fact to react with S. However, while mercury spread its variation also in the other factors, silver shows high EVF for this source. Moreover Ag and Hg are not connected with other heavy metals, suggesting that the hypothesis of an industrial source of pollution could be rejected.

Mercury was in the past used in dental amalgam, together with lower silver and other metal content. However, in factor 2 the main contribution in factor explanation is coming from silver variation.

The high presence of silver could thus be associated with the environmental impact of engineering Ag NPs which flows in municipality due to the high use of this material in house-hold and personal care products. As explained in the chapter introduction, sewage systems are nowadays the main pathway for the release of nanosilver in the environment.

Factor 3

Factor 3 is characterised by the variation of the majority metals and Potassium. Due to the strong variability of sewage samples, being them collected in facilities with different characteristics, the determined source could be explained by a pollution source. This source groups all the metals which could have an anthropogenic influence.

Factor 4

Factor 4 is defined by Fe variation. Since iron (ferrous sulphate) is one of the selected elements used for phosphorus removal at WWTPs facilities, a P-removal source was suggested. In order to have a clearest source identification, G values were explored. It resulted that factor 4 assumes highest values in Finland WWTPs. Since this methodology is widely used in Finland (Ruotsalainen, 2011) we can confirm the factor explanation.

10.3. Conclusions

Monitoring campaign on sewage sludge sample collected at European WWTPs was useful to determine mean values of major, minor and heavy metals content. In addition, comparison with limit values for heavy metals concentration in sludge for their use in agriculture gave satisfactory results.

Moreover, a descriptive statistic and PMF application on inorganic data set allowed drawing conclusions on sludge properties and origins. The first remark was the great variability found in element concentration, evidenced both in boxplots and in factor 3 characterization which grouped, under the same source, all measured metals. In future monitoring campaigns, this problem may be overcome by the selection of more appropriate facilities with common characteristics (i.e.: origin of wastewater, localization, annual load) or increasing their number across Europe).

On the other hand, PMF model reveal a silver-based factor that could be associated with nanosilver content in sewage samples. In order to better understand the factor 1 resolution an the silver-related problem, a further step in the silver factor identification might be the inclusion of organic pollutant originating from domestic wastes (i.e. siloxanes) in the PMF data set. However, data on organic pollutant were not yet completed.

Chapter 11

Conclusions

Basing on results obtained by the positive matrix factorization application, it could be concluded that this statistical approach is a valuable tool for the characterization of different types of environmental data sets, from local to pan-European scales.

Positive matrix factorization well adapted to analyze geochemical data sets, which often contain below-detection-limit data, missing value and outliers, and usually exhibit positively skewed distributions. This property is determined by the use error estimates as individual data weights that allow the algorithm to properly handle these problematic data structures.

The main difference with customary multivariate technique, such as cluster analysis and principal component analysis lies, in fact, that no pre-treatment procedures have to be applied to input data, keeping unchanged the original data structure and prevent loss of information. Results obtained from PCA and PMF comparison, confirm the drawback of PCA to be a data-sensitive method. A careful univariate analysis, acting to detect outliers and remove data skewness and differences in variables range, results in a less accurate sources classification than those estimated by PMF, and often makes PCA interpretation subjective.

The use of outliers as real data and maintaining unchanged positively skewed data structures in PMF resolution, allow extracting as much information as possible from the examined data set. This results, for example, in the identification of a Pb pollution sources in the Alpine lakes application (Ch. 7), and in the characterization of different mineralized components within the *Coren del Cucì* mine site. Moreover, the combination of PMF results with a GIS-based approach confirms an improving on factors characterization, by means of the identification of their impact areas.

For further improvements at the pan-European scale, where different geological and urbanized impacts occur over a large area, sampling location could be selected basing on a common feature; for example, in WWTPs application (Ch. 10) a particular facility types could be selected. Alternatively, the number of samples to collect could be increase across Europe, in order to have a significant number of samples for each country.

In future, PMF could also become a valid tool helping policy-makers to improve/develop environmental policies. Factors identification could lead to the determination of potential marker for contamination sources. Moreover, spatial distribution map of resolved factors can evidence

the role of sub-system (e.g. the role of tributaries in the Danube catchment area). Further monitoring campaign could be planned at the same locations of the examine data set in order to assess changes in the pollution status, i.e. due to a catastrophic event, and consequently revise the regulatory framework.

Finally, it is also important to highlight the importance of method validation in scientific research; it would be a relevant step for the determination of uncertainty estimates to be introduced in the PMF algorithm.

Appendix A: Method validation Data

Appendix A.1 – Precision for low calibration sewage sludge analysis by ICP/AES

	Repeatability	***Between day variation***	***Intermediate precision***
Hg	6%	7%	4%
Ag	6%	2%	6%
Al	-	-	-
As	9%	2%	9%
Ba	3%	4%	5%
Cd	2%	4%	5%
Co	10%	5%	11%
Cr	8%	4%	9%
Cu	3%	1%	3%
Fe	-	-	-
Mg	-	-	-
Mn	3%	2%	4%
Mo	3%	2%	4%
Ni	8%	2%	8%
Pb	9%	4%	10%
Sb	8%	3%	8%
Se	4%	2%	5%
Ti	9%	2%	9%
V	7%	1%	7%
Zn	-	-	-
P	-	-	-
K	-	-	-

Appendix A.2 – Precision for high calibration sewage sludge analysis by ICP/AES

	Repeatability	***Between day variation***	***Intermediate precision***
Hg	10%	11%	10%
Ag	7%	4%	8%
Al	10%	3%	10%
As	1%	3%	3%
Ba	9%	1%	9%
Cd	2%	4%	5%
Co	1%	4%	4%
Cr	1%	1%	1%
Cu	8%	3%	8%
Fe	7%	4%	8%
Mg	4%	6%	7%
Mn	3%	5%	6%
Mo	1%	4%	4%
Ni	1%	2%	2%
Pb	1%	2%	2%
Sb	3%	8%	8%
Se	5%	7%	8%
Ti	1%	8%	8%
V	1%	3%	3%
Zn	5%	3%	6%
P	*1%*	*7%*	*7%*
K	*2%*	*6%*	*6%*

Appendix B: .INI file for PMF2 program

```
##PMF2 .ini file for:  Gromo mine site
## Monitor code M: if M>1, PMF2 writes output every Mth step
## For finding errors, use M<1 to output debug information
##      M       PMF2 version number
       1          4.2
## Dimensions: Rows, Columns, Factors. Number of "Repeats"
                56       12        3        20
##   "FPEAK"  (>0.0 for large values and zeroes on F side)
     0.00000
## Mode(T:robust, F:non-robust)  Outlier-distance        (T=True F=False)
            T                          4.000
## Codes C1 C2 C3 for X_std-dev, Errormodel EM=[-10 ... -14]
       0.0100    0.0000    0.0000     -14
## G Background fit:  Components   Pullup_strength
                           0       0.0000
## Pseudorandom numbers:   Seed     Initially skipped
                             1        0
## Iteration control table for 3 levels of limit repulsion "lims"
##  "lims"    Chi2_test  Ministeps_required  Max_cumul_count
    10.00000   0.50000           5           100
     0.30000   0.50000           5           150
     0.00300   0.30000           5           200
```

Input parameters

```
## Table of FORMATs, with reference numbers from 50 to 59
## Number  Format_text(max 40 chars)
      50   "(A)                                "
      51   "((1X,5G13.5E2))                    "
      52   "((1X,10F8.3))                      "
      53   "((1X,20(I3,:' ')))                 "
      54   "((1X,150(G12.5E1,:' ')))           "
      55   "((1X,180(F9.4,:' ')))              "
      56   "(1X,A)                             "
      57   "((1X,150(G13.5E2,:' ')))           "
      58   "((1X,350(F4.3,:' ')))              "
      59   "((1X,600(I2,:' ')))                "
## Table of file properties, with reference numbers from 30 to 39
## Num- In  Opening  Max-rec File-name(max 40 chars)
## ber  T/F status   length
    30   T "OLD    "  2000  "DATA.txt                                "
    31   T "OLD    "  2000  "T_MAT.txt                               "
    32   T "OLD    "  2000  "V_MAT.txt                                 "
    33   T "OLD    "  2000  "PMF33.DAT                                 "
    34   F "UNKNOWN"  2000  "PMF34.DAT                                 "
    35   F "UNKNOWN"  2000  "PARAMETER_$.TXT"
    36   F "REPLACE"  2000  "G_FACTOR_$.TXT                              "
    37   F "REPLACE"  2000  "F_FACTOR_$.TXT                              "
    38   F "REPLACE"  2000  "TEMP_$.TXT                                  "
    39   F "UNKNOWN"  2000  "$.DAT                                     "
## Input/output definitions for 21 matrices
##  ===HEADING=====   ========MATRIX==========       default HEADING
##   --IN---- --OUT-   -----IN------   ---OUT--       for each matrix
## FIL(R)FMT FIL FMT FIL(R)(C)FMT(T) FIL FMT(T) ------max 40 chars----...
    30 F  50  38  50  30 F      0 F   38  57 F  "X (data matr)             "
    31 F  50  38  56  31 F      0 F   38  57 F  "X_std-dev /T (constant)"
     0 F  50   0  56   0 F      0 F    0  57 F  "X_std-dev /U (sqrt)     "
    32 F  50  38  56  32 F      0 F   38  57 F  "X_std-dev /V (proport) "
     0 F  50   0  56   0 T  F   0 F    0  57 F  "Factor G(orig.)         "
     0 F  50   0  56   0 T  F   0 F    0  57 F  "Factor F(orig.)         "
     0 F  50   0  56   0 F      0 F    0  53 F  "Key (factor G)          "
     0 F  50   0  56   0 F      0 F    0  59 F  "Key (factor F)          "
     0 F  50   0  56   0 F      0 F    0  52 F  "Rotation commands       "
     0 F  50  36  56                  36  57 F  "Computed Factor G Q=    "
     0 F  50  37  56                  37  57 F  "Computed Factor F Q=    "
     0 F  50  36  56                  36  57 F  "Computed std-dev of G  "
     0 F  50  37  56                  37  57 F  "Computed std-dev of F  "
     0 F  50  35  56                  35  57 F  "G_explained_variation  "
     0 F  50  35  56                  35  57 F  "F_explained_variation  "
     0 F  50   0  56                   0  57 F  "Residual matrix X-GF   "
     0 F  50  35  56                  35  57 F  "Scaled resid. (X-GF)/S "
     0 F  50   0  56                   0  57 F  "Robustized residual    "
     0 F  50  35  56                  35  55 F  "Rotation estimates.  Q="
     0 F  50   0  56                   0  55 F  "Computed X_std-dev     "
## If Repeats>1, for input matrices, select (R)=T or (C)=T or none
##    (R)=T: read(generate) again   (C)=T,"chain": use computed G or F
##    none, i.e.(R)=F,(C)=F: use same value as in first task
## (T)=T: Matrix should be read/written in Transposed shape
```

Input and output files

```
## Normalization of factor vectors before output. Select one of:
##   None   MaxG=1   Sum|G|=1 Mean|G|=1  MaxF=1 Sum|F|=1 Mean|F|=1
       T        F        F        F        F        F        F
## Special/read layout for X (and for X_std-dev on following line)
## Values-to-read (0: no special) #-of-X11  incr-to-X12  incr-to-X21
                            0         0         0         0
                            0         0         0         0
## A priori linear constraints for factors, file name: (not yet available)
    "none                                    "
## Optional parameter lines (insert more lines if needed)
    sortfactorsf

## (FIL#4 = this file)    (FIL#24 = .log file)
## After next 2 lines, you may include matrices to be read with FIL=4
## but observe maximum line length = 120 characters in this file
## and maximum line length = 255 characters in the .log fil
```

Optional information

References

Ahamed M., Karns M., Goodson M., Rowe J., Hussain S.M., Schlager J.J., Hong Y., 2008. *DNA damage response to different surface chemistry of silver nanoparticles in mammalian cells*. Toxicology and Applied Pharmacology 233, 404-410.

Arora S., Jain J., Rajwade J.M., Paknikar K.M., 2009. *Interactions of silver nanoparticles with primary mouse fibroblast and liver cells*. Toxicology and Applied Pharmacology 236, 310-318.

Asharani P. V., Wu Y. L., Gong Z., Valiyaveettil S., 2008. *Toxicity of silver nanoparticles in zebrafish models*. Nanotechnology 19, 255102 (8pp).

Anttila P., Paatero P., Tapper U., Järvinen O., 1995. *Source identification of bulk wet deposition in Finland by positive Matrix Factorization*. Atmospheric Environment 14, 1705-1718.

Begum B.A., Kim E., Biswas S.K., Hopke P.K., 2004. *Investigation of sources of atmospheric aerosol at urban and semi-urban areas in Bangladesh*. Atmospheric Environment 38, 3025-3038.

Benn T.M., Westerhoff P., 2008. *Nanoparticles silver released into water from commercially available sock fabrics*. Environmental Science and Technology 42, 4133-4139.

Bhuiyan M.A.H., Parvez L., Islam M.A., Dampare S.B., Suzuki S., 2010. *Heavy metal pollution of coal mine-affected agricultural soils in the northern part of Bangladesh*. Journal of Hazardous Materials 173, 384-392.

Bird G., Brewer P.A., Macklin M.G., Nikolova M., Kotsev T., Mollov M., Swain C., 2010. *Dispersal of Contaminant Metals in the Mining-Affected Danube and Maritsa Drainage Basin, Bulgaria, Eastern Europe*. Water, Air and Soil Pollution 206, 105-127.

Blaser S.A., Scheringer M., MacLeod M., Hungerbühler K., 2008. *Estimation of cumulative aquatic exposure and risk due to silver: contribution of nano-functionalized plastics and textiles*. Science of Total Environment 390, 396-409.

Bostan V., Dominik J., Bostina M., Pardos M., 2000. *Forms of particulate phosphorus in suspension and in bottom sediment in the Danube Delta*. Lakes & Reservoirs: Research and Management 5, 105-110.

Bzdusek P.A., Christensen E.R., Lee C.M., Pakadeesusuk U., Freedman D.C., 2006. *PCB congeners and dechlorination in sediments of Lake Hartwell, South Carolina, determined*

from cores collected in 1987 and 1988. Environmental Science and Technology 40, 109-119.

Chang C.C., Wang J.L., Lung S.C., Liu S.C., Shiu C.J., 2009. *Source characterization of ozone precursors by complementary approaches of vehicular indicator and principal component analysis*. Atmospheric Environment 43, 1771-1778.

Chen X, Schluesener H.J., 2008. *Nanosilver: A nanoproduct in medical application*. Toxicology Letters 176, 1-12.

Chuan M.C., Shu G.Y., Liu J.C., 1996. *Solubility of heavy metals in a contaminated soil: effects of redox potential and pH.* Water, Air, and Soil Pollution 90, 543-556.

Critto A., Carlon C., Marcomini A., 2003. *Characterization of contaminated soil and groundwater surrounding an illegal landfill (S. Giuliano, Venice, Italy) by principal components analysis and kriging*. Environmental Pollution 122, 235-244

Davis J.C., 2002. *Statistics and data analysis in geology*. J.Wiley, New York.

Devesa-Rey R., Díaz-Fierros F., Barral M.T., 2009. *Normalization strategies for river bed sediments: A graphical approach*. Microchemical Journal 91, 253-265.

Dos Santos J.S., De Oliveira E., Bruns R.E., Gennari R.F., 2004. *Evaluation of the salt accumulation process during inundation in water resource of Contas river basin (Bahia–Brazil) applying principal component analysis*. Water Research 38, 1579-1585.

Dragović S., Mihailović N., 2009. *Analysis of mosses and topsoils for detecting sources of heavy metal pollution: multivariate and enrichment factor analysis*. Environmental Monitoring and Assessment 157, 383-390.

Ellison S.L.R., Rosslein M., Williams A., 2000. *EURACHEM/CITAC Guide CG4 - Quantifying uncertainty in analytical measurement – Second edition*. http://www.eurachem.org/guides/pdf/QUAM2000-1.pdf

Fabbricino M., Panico A., Trifuoggi M., 2005. *Copper release in drinking water due to internal corrosion of distribution pipes*. Global NEST Journal 7, 163-171.

Farnham I.M., Singh A.K., Stetzenbach K.J., Johannesson K.H., 2002. *Treatment of nondetects in multivariate analysis of groundwater geochemistry data.* Chemometrics and Intelligent Laboratory Systems 60, 265-281.

Filzmoser P., Garrett R.G., Reimann C., 2005. *Multivariate outlier detection in exploration geochemistry*. Computers and Geosciences 31, 579-587.

Fjällborg B., Dave G., 2003. *Toxicity of copper in sewage sludge*. Environment International 28, 761-769.

Free G., Solimini A.G., Rossaro B., Marziali L., Giacchini R., Paracchini B., Ghiani M., Vaccaro S., Gawlik B., Fresner R., Santner G., Schönhuber M., Cardoso A.C., 2009. *Modelling lake macroinvertebrate species in the shallow sublittoral: relative roles of habitat, lake morphology, aquatic chemistry and sediment composition*. Hydrobiologia 633, 123-136.

Gashi F., Frančišković-Bilinski S., Bilinski H., Troni N., Bacaj M., Jusufi F., 2011. *Establishing of monitoring network on Kosovo Rivers: preliminary measurements on the four main rivers (Drini i Bardhë, Morava e Binqës, Lepenc and Sitnica)*. Environmental Monitoring and Assessment 175, 279-289.

Geranio L., Heuberger M., Nowack B., 2009. *The behaviour of silver nanotextiles during washing*. Environmental Science and Technology 43, 8113-8118.

González A.G., Herrador M.Á., 2007. *A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles*. Trends in analytical chemistry 2 (3).

Gordon G. E., 1988. *Receptor models*. Environmental Science & Technology 22, 1132-1142

Gottschalk F., Sonderer T., Scholz R.W., Nowack B., 2009. *Modelled Environmental concentrations of engineered nanomaterials (TiO_2, ZnO, Ag, CNT, Fullerenes) for different regions*. Environmental Science and Technology 43, 9216-9222.

Grande J.A., Borrego J., De la Torre M.L., Sáinz A., 2003. *Application of cluster analysis to the geochemistry zonation of the estuary waters in the Tinto and Odiel Rivers (Huelva, Spain)*. Environmental Geochemistry and Health 25, 233-246.

Greulich C., Kittler S., Epple M., Muhr G., Köller M., 2009. *Studies on the biocompatibility and the interaction of silver nanoparticles with human mesenchymal stem cells (hMSCs)*. Langenbeck's Archive of Surgery 394, 495-502.

Handy R.D., von der Kammer F., Lead J.R., Hassellöv M., Owen R., Crane M., 2008. *The ecotoxicology and chemistry of manufactured nanoparticles*. Ecotoxicology 17, 287-314.

Hardle W., Simar L., 2003. *Applied Multivariate Statistical Analysis*. TECH method and data technologies.

Helstrup T., Jørgensen N.O., Banoeng-Yakubo B., 2007. *Investigation of hydrochemical characteristics of groudwater from the Cretaceous-Eocene limestone aquifer in southern*

Ghana and souther Togo using hierarchical cluster analysis. Hydrogeology Journal 15, 977-989.

Hoffmann G., Schingnitz D., Bilitewski B., 2010. *Comparing different methods of analysing sewage sludge, dewatered sewage sludge and sewage sludge ash*. Desalination 250, 399–403

Houhou J., Lartiges B.S., Montarges-Pelletier E., Sieliechi J., Ghanbaja J., Kohler A., 2009. *Sources, nature, and fate of heavy metal-bearing particles in the sewer system*. Science of the Total Environment 407 (2009) 6052–6062

Huang S., Rahn K.A., Arimoto R,. 1999. *Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island*. Atmospheric Environment 33, 2169-2185.

Huang S., Conte M.H., 2009. *Source/process apportionment of major and trace elements in sinking particles in the Sargasso sea*. Geochimica et Cosmochimica Acta 73, 65-90.

ICPDR – International Commission for the Protection of the Danube River, 2008. *Joint Danube Survey 2, Final Scientific Report*. ICPDR Secretariat, Vienna International Centre, D0412, P.O. Box 500, 1400 Vienna, Austria.

ISO, 1995. *Soil quality. Extraction of Trace Elements Soluble in Aqua Regia.*

ISO Guide 35, 1989. *Certification of reference materials – general and statistical principles*. International Organization for Standardization, CH-1211 Genève 20, Switzerland

ISO/IEC Guide 98-3, 2008. *Guide to the expression of uncertainty in measurement (GUM)*.

Jervis G., 1881. *I tesori sotterranei d'Italia. Parte prima: la regione delle Alpi*. Loescher, Torino.

Jolliffe I.T., 2002. *Principal Component Analysis – Second edition*. Springer.

Juntto S., Paatero P., 1994. *Analysis of daily precipitation data by positive matrix factorization*. Environmetrics 5, 127-144

Kabata-Pendias A., Pendias H., 2001. *Trace elements in soils and plants*. CRC, Boca Raton, pp. 413.

Kaplunovsky A.S., 2005. *Factor analysis in environmental studies*. HAIT Journal of Science and Engineering 2 54-94.

Keshav Krishna A., Rama Mohan K., Murthy N.N., 2011. *A multivariate statistical approach for monitoring of heavy metals in sediments: a case study from Wailpalli*

Watershed, Nalgonda District, Andhra Pradesh, India. Research Journal of Environmental and Earth Sciences 3, 103-113.

Kim B., Park C-S., Murayama M., Hochella M.F., 2010. *Discovery and characterization of silver sulfide nanoparticles in final sewage sludge products*. Environmental Science and Technology 44, 7509-7514.

Kirschner A.K.T., Kavka G.G., Velimirov B., Mach R.L., Sommer R., Farnleitner A.H., 2009. *Microbiological water quality along the Danube River: Integrating data from two whole-river surveys and a transnational monitoring network.* Water Research 43, 3673-3684.

Klaver, G., van Os, B., Negrel, P., Petelet-Giraud, E., 2007. *Influence of hydropower dams on the composition of the suspended and riverbank sediments in the Danube*. Environmental Pollution 148, 718-728.

Kvitek L., Vanickova M., Panacek A., Soukupova J., Dittrich M., Valentova E., Prucek R., Bancirova M., Milde D., Zboril R, 2008. *Initial study on the toxicity of silver nanoparticles (NPs) against Paramecium caudatum*. The Journal of Physical Chemistry 113, 4296-4300.

Laurin D., Stupar J., June 1987. *Antimicrobial compositions*. U.S. patent 4,677,143.

Lee E., Chan C.K., Paatero P., 1999. *Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong.* Atmospheric Environment 33, 3201-3212.

Lingwall J., Christensen W.F., 2007. *Pollution source apportionment using a priori information and positive matrix factorization*. Chemometrics and Intelligent Laboratory Systems 87, 281-294.

Literathy P., Laszlo F., 1995. *Harmonization of micropollutant monitoring in large international river: Danube.* Water Science and Technology 32, 125-137.

Loco J.V., Elskens M., Croux C., Beernaert H., 2002. *Linearity of calibration curves: use and misuse of the correlation coefficient.* Accreditation and Quality Assurance 7, 281-285.

Loska K., Wiechuła D., 2003. *Application of principle component analysis for the estimation of source of heavy metal contamination in surface sediments from the Rybnik Reservoir*. Chemosphere 51, 723-733.

Lu J., Jiang P., Wu L., Chang A.C., 2008. *Assessing soil quality data by positive matrix factorization*. Geoderma 145, 259-266.

Melas D., Zerefos C., Rapsomanikis S., Tsangas N., Alexandropoulou A., 2000. *The war in Kosovo-Evidence of pollution transport in the Balkans during operation "Allied Force".* Environmental Science and Pollution Research 7, 97-104.

Milačič R., Ščančar J., Murko S., Kocman D., Horvat M., 2010. *A complex investigation of the extent of pollution in sediments of the Sava River. Part 1: Selected elements.* Environmental Monitoring and Assessment 163, 263-275.

Mitchell R.L., 1964. In: Beer F.E. - *Chemistry of the soils*, p.515.Van Nostrand Reinhold, New York.

Morones J. R., Elechiguerra J. L., Camacho A., Holt K., Kouri J. B., Ramirez J. T., Yacaman M. J., 2005. *The bactericidal effect of silver nanoparticles*. Nanotechnology 16, 2346-2353

Morrison J.M., Goldhaber M.B., Ellefsen K.J., Mills C.T., 2011. *Cluster analysis of a regional-scale soil geochemical dataset in northern California*. Applied Geochemistry 26, 105-107.

Mostert M.M.R., Ayoko G.A., Kokot S., 2010. *Application of chemometrics to analysis of soil pollutants.* Trends in Analytical Chemistry 29, 430-445.

Motelay-Massei A., Ollivon D., Garban B., Chevreuil M., 2003. *Polycyclic aromatic hydrocarbons in bulk deposition at a suburban site: assessment by principal component analysis of the influence of meteorological parameters*. Atmospheric Environment 37, 3135-3146.

Mueller N.C., Nowack B., 2008. *Exposure modeling of engineered nanoparticles in the environment*. Environmental Science and Technology 42, 4447-4453.

Norris G., Vedantham R., Wade K., Brown S., Prouty J., Foley C., 2008. *EPA Positive Matrix Factorization (PMF) 3.0: fundamentals & user guide*. U.S. Environmental Protection Agency

Officer S.J., Kravchenko A., Bollero G.A., Sudduth K.A., Kitchen N.R., Wieboid W.J., Palm H.L., Bullock D.G., 2004. *Relationships between soil bulk electrical conductivity and the principal component analysis of topography and soil fertility values*. Plant Soil 258, 269-280.

Ogulei D., Hopke P.K., Wallace L.A., 2006. *Analysis of indoor particle size distributions in an occupied townhouse using positive matrix factorization*. Indoor air 16, 204-215.

Paatero P., 1997. *Least square formulation of robust non-negative factor analysis*. Chemometrics and Intelligent Laboratory Systems 37, 23-35.

Paatero P., 1999. *The multilinear engine–a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model*. Journal of Computational and Graphical Statistics 8, 854-888.

Paatero P., 2007a. *User's guide for Positive Matrix Factorization programs PMF2 and PMF3, Part1: tutorial.* University of Helsinki, Helsinki, Finland.

Paatero P., 2007b. *User's guide for Positive Matrix Factorization programs PMF2 and PMF3, Part2: references.* University of Helsinki, Helsinki, Finland.

Paatero P., 2007c. *End User's Guide to Multilinear Engine Applications*. ftp://ftp.clarkson.edu/pub/hopkepk/pmf/.

Paatero P., Hopke P.K., 2003. *Discarding or downweighting high-noise variables in factor analytic models*. Analytica Chimica Acta 490, 277-289.

Paatero P., Hopke P.K., Begum B.A., Biswas S.K., 2005. *A graphical diagnostic method for assessing the rotation in factor analytical models of atmospheric pollution*. Atmospheric Environment 39, 193-201.

Paatero P., Hopke P.K., Song X.-H., Ramadan Z., 2002. *Understanding and controlling rotations in factor analytic models*. Chemometrics and Intelligent Laboratory Systems 60, 253-264.

Paatero P., Tapper U., 1993. *Analysis of different modes of factor analysis as least squares fit problems*. Chemometrics and Intelligent Laboratory System 18, 183-194.

Paatero P., Tapper U., 1994. *Positive Matrix Factorization: a non-negative factor model with optimal utilization of error estimates of data values*. Environmetrics 5, 111-126.

Paunović M., Vassile V., Cheshmedjie S., Simić V., 2008. *Environmental and Risk Assessment of the Timok River Basin*. Regional Environmental Center.

Pawellek F., Frauenstein F., Veizer J., 2002. *Hydrochemistry and isotope geochemistry of the upper Danube River.* Geochimica et Cosmochimica Acta 66, 3839-3854.

Perelshtein I., Applerot G., Perkas N., Guibert G., Mikhailov S., Gedanken A., 2008. *Sonochemical coating of silver nanoparticles on textile fabrics (nylon, polyester and cotton) and their antibacterial activity*. Nanotechnology 19, 245705 (6pp).

Pires J.C.M., Sousa S.I.V., Pereira M.C., Alvim-Ferraz M.C.M., Martins F.G., 2008. *Management of air quality monitoring using principal component and cluster analysis – part I: SO2 and PM10*. Atmospheric Environment 42, 1249–1260.

Pizarro J., Vergara P.M., Rodríguez J.A., **Valenzuela A.M**., 2010. *Heavy metals in northern Chilean rivers: Spatial variation and temporal trends*. Journal of Hazardous Materials 181, 747-754.

Polissar A.V., Hopke P.K., Malm W.C., Sisler J.F., 1998. *Atmospheric aerosol over Alaska: 2. Elemental composition and sources*. Journal of Geophysical Research 103, 19045-19057.

Polissar A.V., Hopke P.K., Paatero P., Kaufmann Y.J., Hall D.K., Bodhaine B.A., Dutton E.G., Harris J.M., 1999. *The aerosol at Barrow, Alaska: long-term trends and source locations*. Atmospheric Environment 33, 2441-2458.

Polissar A.V., Hopke P.K., Poirot R.L., 2001. *Atmospheric aerosol over Vermont: chemical composition and sources*. Environmental Science & Technology 35, 4604-4621.

Pollice A., 2009. *Recent statistical issues in multivariate receptor models*. Environmetrics 22, 35-41.

Preacher K.J., MacCallum R.C., 2003. *Repairing Tom Swift's electric factor analysis machine*. Understanding Statistics 2, 13-43.

Puura E., Marmo L., D'Alessandro M., 2002. *Workshop on mine and quarry waste-the burden from the past*. European Commission – Joint Research Centre, EUR 20661 EN.

R Development Core Team, 2005. *R: A language and environment for statistical computing, reference index version 2.12.1.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Reff A., Eberly S.I., Bhave P.V., 2007. *Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods*. Journal of the Air & Waste Management Association 57, 146-154

Reid M.K., Spencer K.L., 2009. *Use of principal component analysis (PCA) on estuarine sediment datasets: The effect of data pre-treatment*. Environmental Pollition 157, 2275-2281.

Reimann C., Filzmoser P., Garrett R.G., 2002. *Factor analysis applied to regional geochemical data: problems and possibilities.* Applied geochemistry 17, 185-206.

Reinikainen S.-P., Laine P., Minkkinen P., Paatero P., 2001. *Factor analytical study on water quality in Lake Saimaa, Finland*. Fresenius' Journal of Analytical Chemistry 369, 727-732.

Relić D., Đorđević D., Popović A., Blagojević T., 2005. *Speciations of trace metals in the Danube alluvial sediments within an oil refinery.* Environment International 31, 661-669.

Ribeiro J., Ferreira da Silva E., Li Z., Ward C., Flores D., 2010. *Petrographic, mineralogical and geochemical characterization of the Serrinha coal waste pile (Douro Coalfield, Portugal) and the potential environmental impacts on soil, sediments and surface waters*. International Journal of Coal Geology 83, 456-466.

Ruotsalainen I., 2011. *Waste water treatment plants in the baltib sea drainage basin.* Master thesis, Aalto University - School of Engineering, Department of Civil and Environmental Engineering.

Sakan S.M., Đorđević D.S., Manijlović D.D., Predrag P.S., 2009. *Assessment of heavy metal pollutants accumulation in the Tisza river sediments.* Journal of Environmental Management 90, 3382-3390.

Sakan S.M., Đorđević D.S., Manijlović D.D., 2010. *Trace elements as tracers of environmental pollution in the canal sediments (alluvial formation of the Danube River, Serbia)*. Environmental Monitoring and Assessment 167, 219-233.

Salminen R., Tarvainen T., Demetriades A., Duris M., Fordyce F.M., Gregorauskiene V., Kahelin H., Kivisilla J., Klaver G., Klein H., Larson J.O., Lis J., Locutura J., Marsina K., Mjartanova H., Mouvet C., O'Connor P., Odor L., Ottonello G., Paukola T., Plant J.A., Reimann C., Schermann O., Siewers U., Steenfelt A., Van der Sluys J., deVivo B., Williams L., 1998. *FOREGS Geochemical mapping field manual.* Geological Survey of Finland, Guide 47.

Santos Bermejo J.C., Beltrán R., Gómez Ariza J.L., 2003. *Spatial variations of heavy metals contamination in sediments from Odiel river (Southwest Spain)*. Environment International 29, 69-77

Schaefer K., Einax J.W., Simeonov V., Tsakovski S., 2010. *Geostatistical and multivariate statistical analysis of heavily and manifoldly contaminated soil samples*. Analytical and Bioanalytical Chemistry 396, 2675-2683.

Servida D., De Capitani L., Grieco G., 2006. *Caratterizzazione geochimica e biogeochimica della discarica mineraria del Coren del Cucì di Gromo (Alta Val Seriana, BG).* 85° Congresso SIMP, Fluminimaggiore.

Servida D., Moroni M., Ravagnani D., Rodeghiero F., Venerandi I., De Capitani L., Grieco G., 2010. *Phreatic sulphide bearing quartz breccias between crystalline basement and Collio formation (Southern Alps, Italy)*. Italian Journal of Geosciences 129, 223-236.

Sheng Y., Fu G., Chen F., Chen J., 2011. *Geochemical characteristics of inorganic sulfur in Shijing River, South China*. Journal of Environmental Monitoring 13, 807-812.

Swanson S.K., Bahr J.M., Schwar M.T., Potter K.W., 2001. *Two-way cluster analysis of geochemical data to constrain spring source waters*. Chemical Geology 179, 73-91.

Tauler R., Peré-Trepat E., Lacorte S., Barceló D., 2004. *Chemometrics Modeling of Environmental Data*. Department of Environmental Chemistry, Institute of Chemical and Environmental Research IIQAB-CSIC, Spain.

Templ M., Filzmoser P., Reimann C., 2008. *Cluster analysis applied to regional geochemical data: Problems and possibilities*. Applied Geochemistry 23, 2198- 2213.

Treffeisen R., Herber A., Ström J., Shiobara M., Yamanouchi T., Yamagata S., Holmén K., Krievs M., Schrems O., 2004. *Interpretation of Artic aerosol properties using cluster analysis applied to observations in the Svalbard area*. Tellus 56B, 457-476.

Unonius L., Paatero P., 1990. *Use of singular value decomposition for analyzing repetitive measurements*. Computer Physics Communications 59, 225-243

United Nation Economic Commission for Europe (UNECE), 2007. *Our waters: Joining hands across borders. First Assessment of Transboundary Rivers, Lakes and Groundwaters*. United Nation, New York and Geneva.

U.S. EPA - United States Environmental Protection Agency, 2000. *Abandoned mine site characterization and cleanup handbook* EPA 910-B-00-001, Seattle.

U.S. EPA - United States Environmental Protection Agency, 2009. *Targeted National Sewage Sludge Survey Sampling and Analysis Technical Report.* http://water.epa.gov/scitech/wastetech/biosolids/upload/2009_01_15_biosolids_tnssss-tecg.pds

Vaccaro S., Sobiecka E., Contini S., Locoro G., Free G., Gawlik B.M., 2007. *The application of positive matrix factorization in the analysis, characterization and detection of contaminated soils*. Chemosphere 69, 1055-1063.

Viana M., Kuhlbusch T.A.J., Querol X., Alastuey A., Harrison R.M., Hopke P.K., Winiwarter W., Vallius M., Szidat S., Prévôt A.S.H., Hueglin C., Bloemen H., Wåhlin P., Vecchi R., Miranda A.I., Kasper-Giebl A., Maenhaut W., Hitzenberger R., 2008. *Source apportionment of particulate matter in Europe: A review of methods and results*. Aerosol Science 39, 827-849

Vogel B., Pall K., 2002. *Chapter 3: Nine Geo-morphological Danube Reaches. Joint Danube Survey (JDS). Technical Report of the International Commission for the Protection of*

the Danube River. In: Literáthy, P., V. Koller-Kreimel, and I. Liška, 2002. Joint Danube Survey. Technical Report of the International Commission for the Protection of the Danube River, 22-31.

Webster R., 2001. *Statistic to support soil research and their presentation*. European Journal of Soil Science 52, 331-340.

Woitke P., Wellmitz J., Helm D., Kube P., Lepom P., Litheraty P., 2003. *Analysis and assessment of heavy metal pollution in suspended solids and sediments of the river Danube.* Chemosphere 51, 633-642.

Woodrow Wilson International Center for Scholars, 2009. *A Nanotechnology Consumer Products Inventory.* www.nanotechproject.org/consumerproducts.

Xie Y., Berkowitz C.M., 2006. *The use of positive matrix factorization with conditional probability functions in air quality studies: an application to hydrocarbon emissions in Houston, Texas*. Atmospheric Environment 40, 3070-3091.

Yiğiterhan O., Murray J.W., 2008. *Trace metal composition of particulate matter of the Danube River and Turkish rivers draining into the Black Sea*. Marine Chemistry 111, 63-76.

Yongming H., Peixuan D., Junji C., Posmentier E.S., 2006. *Multivariate analysis of heavy metal contamination in urban dusts of Xi'an, central China*. Science of the Total Environment 355, 176-186.

Yu T.Y., Chang L.F.W., 2000. *Selection of the scenarios of ozone pollution at southern Taiwan area utilizing principal component analysis*. Atmospheric Environment 34, 4499-4509.

Zaharescu D.G., Hooda P.S., Soler A.P., Fernandez J., Burghelea C.I., 2009. *Trace metals and their source in the catchment of the high altitude Lake Respomuso, Central Pyrenees*. Science of Total Environm

Zupan M., Einax J.W., Kraft J., Lobnik F., Hudnik V., 2000. *Chemometric characterization of soil and plant pollution. Part 1: multivariate data analysis and geostatistical determination of relationship and spatial structure of inorganic contaminants in soil*. Environmental Science and Pollution Research 7, 89-96.

Acknowledgements

A volte i ringraziamenti sembrano la parte più complicata da scrivere perché ci si fanno mille domande su chi includere e chi no: colleghi? Amici? Famiglia?
Mi ero sentita quasi obbligata a doverli scrivere e così ho preferito ringraziare a voce tutte le persone che mi hanno aiutato durante questi tre anni di dottorato, sia lavorativamente che affettivamente. Tuttavia mi è sembrato giusto lasciare impronta di chi ha partecipato attivamente alla realizzazione del progetto.
Ai miei relatori, la Prof.ssa De Capitani e il Prof. Gawlik per avermi lasciato tutta l'indipendenza possibile;
ai miei colleghi Giovanni, Carmen, Isabelle e Agustin per il loro aiuto in laboratorio;
a Diego e Stefano per i loro preziosi input e suggerimenti;
e a Kevin per il suo aiuto con l'inglese.

Grazie!