S. Afr. J. Chem., 2003, **56**, 15–20,
<http://journals.sabinet.co.za/sajchem/>.

Analysis of Trace Elements in South African Clinkers using Latent Variable Model and Clustering

János Abonyi[a], Ferenc D. Tamás[b], Sanja Potgieter[c*] and Herman Potgieter[d]

[a]*Department of Process Engineering, University of Veszprém, P.O. Box 158, Veszprém, H-8201 Hungary.*
[b]*Department of Silicate- and Materials Engineering, University of Veszprém, P.O. Box 158, Veszprém, H-8201 Hungary.*
[c]*Department of Chemistry and Physics, Technikon Pretoria, Private Bag X680, Pretoria, 0001 South Africa.*
[d]*Department of Chemical and Metallurgical Engineering, Technikon Pretoria, Private Bag X680, Pretoria, 0001 South Africa.*

Received 12 June 2002; revised 18 March 2003; accepted 19 March 2003

ABSTRACT

The trace element content of clinkers (and possibly of cements) can be used to identify the manufacturing factory. The Mg, Sr, Ba, Mn, Ti, Zr, Zn and V content of clinkers give detailed information for the determination of the origin of clinkers produced in different factories. However, for the analysis of such complex data there is a need for algorithmic tools for the visualization and clustering of the samples. This paper proposes a new approach for this purpose. The analytical data are transformed into a two-dimensional latent space by factor analysis (probabilistic principal component analysis) and dendograms are constructed for cluster formation. The classification of South African clinkers is used as an illustrative example for the approach.

KEY WORDS

Clinker, trace elements, factor analysis, principal component analysis, clustering, dendogram.

1. Introduction

The trace element content of clinkers is of significant scientific interest, and can be used to solve practical problems too, e.g. to determine the origin of the clinker (i.e. the manufacturing works). The first paper on a similar topic was published in 1993 by Goguel and St John,[1] and showed the Ba, Sr and Mn concentration of different Portland cements in New Zealand concretes. This first attempt suggests that advanced statistical methods, so-called 'pattern recognition' or 'fingerprinting', can help with qualitative identification.[2]

However, the qualitative identification obviously requires a database to compare the trace element content of unknown clinkers/cements with characteristic known samples. Data describing trace element content of clinkers and cements have been published previously.[3,4,5] In these papers it was shown that not all trace elements could be used for fingerprinting; the selection of particular elements for identification purposes must follow certain principles. The most important criteria of selection is that trace elements of 'dactylogrammatic value' should come from the main raw materials (limestone, marl, clay) and not from the fuel, the furnace lining or from grinding media wear. Some other principles should be observed as well. More recently, six elements were used to characterize clinkers: besides those used by Goguel and St John,[1] the Mg, Ti and Zr contents of clinkers were also employed.[4,5] Zn and V have no dactylogrammatic value because they come from the fuel, e.g. when waste tyres or special sorts of heavy fuel oil are used, respectively.

In a previous paper[6] the dactylogrammatic value of trace elements was described, together with detailed data on sample preparation, analysis, averages and standard deviations of eight trace elements (Mg, Sr, Ba, Mn, Ti, Zr, Zn and V). Based on more than 200 samples, a 'standard' trace element content was calculated and, in order to facilitate the visualization of the trace element content, a graphic method ('Star Plotting') was presented, where every clinker is compared to the proposed standard.

This paper explored the next step in the presentation of the data analysis of the trace element content of clinkers, by trying to establish whether a useful structure based on distinct simple groups could be discerned. Furthermore, if a clinker sample could be classified into one of these groups would it be possible to predict some of its properties from it. The first issue is addressed using principal component analysis (PCA) or cluster analysis, whereas the second one is investigated through the use of pattern recognition methods.[7]

PCA is the most widely used multivariate analysis technique in science and engineering. It is a method for transforming the original measurement variables into new variables called principal components. By plotting the data in a coordinate system defined by the two or three largest principal components it is possible to identify the key relationships in the data, that is, find similarities and differences among objects (such as different clinkers) in a data set. In previous papers the analytical data were transformed by principal component analysis and dendograms were constructed for cluster formation.[3,4,5]

A persistent weakness of PCA in chemical applications has been its inability to handle measurement uncertainty. In a previous paper[8] a new technique was described which is referred to as 'maximum likelihood principal component analysis' because of the incorporation of measurement variance information in the model estimation. Unfortunately, the successful implementation of this algorithm requires estimate of the error covariance and this hampers the application of the method. Furthermore, the iterative algorithm requires much computation, e.g. results obtained in previous work[9] required computational times ranging from one hour to more than a day.

Tipping & Bishop have developed another approach for the determination of principal axes of a set of observed data vectors through maximum-likelihood estimation of parameters in a latent variable model that is closely related to factor analysis.[10]

* To whom correspondence should be addressed; E-mail: potgieters@techpta.ac.za

Table 1 Trace element content of South African clinkers (mg/kg).

Code	Ba	Mn	Sr	Ti	Zr	Mg	V	Zn	Factory
SA 1	146	428	1024	853	18	2693	20	9	1
SA 10	155	451	1192	943	30	3165	20	9	1
SA 16	168	444	1168	893	0	3298	22	19	1
SA 23	235	386	2110	1061	54	5241	23	106	1
SA 2	569	3003	49	1178	32	15265	47	39	2
SA 9	604	2933	19	1242	63	15838	44	13	2
SA 3	485	7020	213	1052	41	24292	29	12	3
SA 8	558	6566	179	1126	73	23225	26	11	3
SA 17	407	6638	164	1136	33	24125	26	18	3
SA 20	449	4519	168	957	68	20973	29	29	3
SA 4	207	584	2090	25	30	5356	27	26	4
SA 5	210	610	2126	1176	9	5358	26	25	4
SA 12	195	509	2296	1224	18	5659	24	26	4
SA 13	193	490	2298	1252	45	5420	24	28	4
SA 15	191	497	2274	1208	32	5523	24	30	4
SA 21	176	379	2142	1102	56	5039	23	29	4
SA 22	174	434	1058	880	50	3126	19	21	4
SA 6	122	264	2934	804	15	5680	17	40	5
SA 11	136	210	3107	903	47	6723	19	14	5
SA 19	165	491	3484	805	31	6314	20	40	5

This technique has been successfully applied to many problems in computer science[11], but its chemometrics relevant applications have been not studied yet. The aim of this paper is to show how this tool can be effectively used in the problem of the qualitative identification of clinkers produced in different factories.

In the next section of this paper a short description of the project launched for the collection and chemical analysis of clinkers is given. Then the theoretical background of the data visualization (projection) and clustering tools is presented and in the final part, before the conclusions, the results of a factual example of the clustering and classification of South African clinkers are given. This example illustrates that the proposed method is useful to visualize the samples, and to identify compact models that are able to determine the origin of the clinker.

2. Experimental

2.1 Materials

For the qualitative 'fingerprinting' of clinkers, obviously a set of well-defined clinker samples is necessary. To obtain such an informative database, a Technical Committee (TC 180/QIC) (Qualitative Identification of Clinkers and Cements) was established in 1996, under the auspices of RILEM (Réunion Internationale des Laboratoires d'Essais et de Recherches sur les Matériaux et les Constructions). This project was aimed at the collection of composite average samples from eight countries (Austria, Portugal, South Africa, Slovakia, Slovenia, Spain, Switzerland and the United Kingdom). Over 200 samples were collected and analysed. This paper focuses only on the South African clinkers collected during the project. Approximately twenty clinker samples have been collected from five South African cement factories and their Mg, Sr, Ba, Mn, Ti, Zr, Zn and V contents determined. The results of the chemical analysis are given in Table 1.

2.2. Methods

The mass of each clinker sample that arrived at the Veszprém laboratory (Hungary) was approx. 2–3 kg. It was usually in the form of uncrushed nodules. This was then crushed and a smaller average sample was taken according to sampling standards. This smaller amount was ground in a centrifugal mill. The final size reduction (smaller than a sieve size of 63 μm) was done by hand in an agate mortar. A preliminary experiment with pure quartz showed that the abrasion of grinding media or mill lining did not cause any significant pollution of the sample for the elements analysed. An exactly weighed sample (ca. 1 g) was dissolved in hydrochloric acid. The precipitated SiO_2 was filtered off, washed and the filtrate analysed by ICP-ES (Inductively Coupled Plasma Optical Emission Spectrography). Duplicate samples were prepared of all clinkers for analysis, and if the difference was >10%, the sample preparation and analysis was repeated.

Initially a ARL-3410 ICP employing a 27 MHz radio frequency at a power of 650 W was used. It was later replaced with a GBC-Integra-XM-type ICP spectrometer, which was equipped with a mini-plasma torch and had a 40 MHz radio-frequency generator of 2 kW power. The spectral range investigated covered 165–800 nm. Computations were performed using EPIC (Evolutionary Program for Instrument Control) software on an IBM PS/2 computer. The wavelengths used in this study were (in nm): Mn = 257.610, Mg = 279.553, Sr = 407.771, Ba = 455.403, Ti = 336.121, Zr = 349.621, Zn = 213,856 V = 310.230.

The measured trace element contents of the South African clinkers are shown in Table 1.

3. Probabilistic PCA and Clustering

3.1. Probabilistic PCA for Factor Analysis

In our research work, the visualization and the clustering of the trace element content of clinkers is considered. Hence, the data are the measurements (observations) of the trace element contents of the clinkers. Each observation consists of n measured variables, grouped into an n-dimensional column vector $\mathbf{x}_k = [x_{1k,\ldots,x_{nk}}]^T, \mathbf{x}_k \in \Re^n$. A set of N observations is denoted by $\mathbf{X}=[\mathbf{z}_1,\ldots,\mathbf{z}_N]^T$ and represented as a $N \times n$ matrix. In pattern recognition terminology, the rows of $\mathbf{X}$ are called patterns or objects, the columns are called the features or attributes, and $\mathbf{X}$ is called the pattern matrix.

The q principal components of the observed data vector $\mathbf{x}_k$ are given by the vector $\mathbf{z}_k=\mathbf{W}^T(\mathbf{x}^k-\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ represents the mean of the data, and $\mathbf{W}^T=[\mathbf{w}_1,K,\mathbf{w}_q]^T$ the transformation matrix, where $\mathbf{w}_j$ $j=1,K,q$ are principal component axes. It can be shown that these ortho-normal axes are given by the q eigenvectors of the sample covariance matrix $\mathbf{S}=1/N\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^T$ such that $\mathbf{S}\mathbf{w}_j=\lambda_j\mathbf{w}_j$.

A latent variable mode seeks to relate the set of n-dimensional observed data vector, $\mathbf{x}_k$, to a corresponding set of q-dimensional latent variables, $\mathbf{z}_k$, $\mathbf{x}=\mathbf{W}\mathbf{z}+\boldsymbol{\mu}+\boldsymbol{\varepsilon}$ where the latent variables have a unit isotropic Gaussian distribution, $\mathbf{z}\approx N(\mathbf{0},\mathbf{I})$

$$p(\mathbf{z})=(2\pi)^{-q/2}\exp\left(-\left\|\mathbf{z}^T\mathbf{z}\right\|^2\right).$$

The error, or noise model is also Gaussian, $\boldsymbol{\varepsilon}\approx N(\mathbf{0},\boldsymbol{\Psi})$ with diagonal $\boldsymbol{\Psi}$. For the case of isotropic noise, $\boldsymbol{\varepsilon}\approx N(\mathbf{0},\sigma^2\mathbf{I})$, the probability distribution over the data space for a given $\mathbf{z}$ is

$$\mathrm{p}(\mathbf{x}\,|\,\mathbf{z})=\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{x}-\mathbf{W}\mathbf{z}+\boldsymbol{\mu}\right\|^2\right).$$

Given this formulation, the model for $\mathbf{x}$ is also a normal distribution $\mathbf{x}=N(\boldsymbol{\mu}\mathbf{C})$, where the $\mathbf{C}$ variance is $\mathbf{C}=\boldsymbol{\Psi}+\mathbf{W}\mathbf{W}^T$,

$$p(\mathbf{x})=(2\pi)^{-n/2}\left|\mathbf{C}\right|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Using Bayes' rule, the posterior distribution of the latent variable $\mathbf{z}$ given by the observed $\mathbf{x}$ may be calculated:

$$p(\mathbf{z}\,|\,\mathbf{x})=(2\pi)^{-q/2}\left|\sigma^{-2}\mathbf{M}\right|^{-1/2}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\right)^T\left(\sigma^{-2}\mathbf{M}\right)\left(\mathbf{x}-\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu})\right)\right)$$

where $\mathbf{M}=\sigma^2\mathbf{I}+\mathbf{W}^T\mathbf{W}$.

Thus the intention is that the dependencies between the data variables $\mathbf{x}$ are explained by a smaller number of latent variables $\mathbf{z}$ while $\boldsymbol{\varepsilon}$ represents the independent noise. This is in contrast with PCA that treats the inter-variable dependecies and the independent noise identically. In factor analysis the columns of $\mathbf{W}$ will generally not correspond to the principal subspace of the data and their values must be determined together with $\boldsymbol{\Psi}$ by maximizing the log-likelihood of the data. In was previously shown[10] that the solution of $\mathbf{W}$ is

$$\mathbf{W}=\mathbf{U}_q\left(\boldsymbol{\Lambda}_q-\sigma^2\mathbf{I}\right)^{1/2}\mathbf{R},$$

where the q column vectors in $\mathbf{U}_q$ are the eigenvectors of $\mathbf{S}$, with corresponding eigenvalues in the $\boldsymbol{\Lambda}_q$ diagonal matrix, and $\mathbf{R}$ is an arbitrary $q\times q$ rotation matrix. Note that because of the use of the $\mathbf{W}^T\mathbf{W}$ term, the likelihood (and the model) is invariant with respect to $\mathbf{R}$.

The variance of the model can be calculated and interpreted as the variance lost in the projection averaged over the lost dimensions

$$\sigma^2=\frac{1}{n-q}\sum_{j=q+1}^{n}\lambda_j.$$

3.2. Hierarchical Clustering of the Latent Variables

Plotting the data in a coordinate system defined by the first two or three columns of $\mathbf{W}$ often provides more than enough information about the overall structure of the data. However, for the automatic detection of groups of the objects (clinkers) the application of clustering algorithm can be used, where clustering attempts to find clusters of patterns (i.e data points) in the latent space. Although several clustering algorithms exist, e.g. K-means, Fuzzy C-variaties[12], hierarchical clustering is by far the most widely used clustering method[7]. The starting point for a hierarchical clustering experiment is the similarity matrix which is formed by first computing the distances between all pairs of points in the data set. The distance in space between the points is determined by some distance function. In this case the Euclidian distance function was used, whereby $ED_{ik}=\left[\sum_{j=1}^{n}(x_{i,j}-x_{k,j})^2\right]^{1/2}$, where $x_{i,j}$ and $x_{k,j}$ are the measured values of the jth parameter of object i and k and n = number of parameters of space dimensions. The similarity value is given by $S_{ik}=1-\frac{ED_{i,k}}{ED_{\max.}}$ and gives a measure of the similarity between objects. The similarity values are organized in the form of a table or matrix. The similarity matrix is then scanned for the largest value, which corresponds to the most similar data pair. The two samples constituting the pair are combined to form a new point, which is located midway between the two original points. The rows and columns corresponding to the old data points are then removed from the matrix. The similarity matrix for the data set is then recomputed. This process is repeated until all points have been linked. There are a variety of ways to compute the distances between data points and clusters in hierarchical clustering. The utilized single-linkage method assesses similarity by measuring the distance to the farthest point in the cluster.

The results of a hierarchical clustering are usually displayed as a dendogram, which is a tree-shaped map of the inter-sample distances in the data set. The dendogram shows the merging of samples into clusters at various stages of the analysis and the similarities at which the clusters merge, which the clustering displayed hierarchically. Interpretation of the results is intuitive, which is the major foundation of these methods.

4. Results

Hierarchical clustering methods attempt to uncover the intrinsic structure of a multivariate data set without making prior assumption about the data. Hence during the identification of the model the class labels (factories, last column of Table 1) were assumed to be unknown. The presented probabilistic latent variable model and the hierarchical clustering method have been implemented in MATLAB©. The program can be downloaded from http://www.fmt.vein.hu/softcomp.

As an illustration of the complexity of the problem, the eight dimensional space of the trace element content has been projected to all of the possible combinations of the trace elements in Fig. 1, where every rows and columns represent one trace element, in such a way that the first subfigure in each row defines the histogram of the given trace element.

To illustrate the advantages of the probabilistic PCA model, three different models were identified:

Model 1. Hierarchical clustering of the observed data

Model 2. PCA projection of the data into a two-dimensional space and hierarchical clustering of the projected data

Model 3. Application of two-dimensional probabilistic PCA, and hierarchical clustering of the latent variables

4.1. Model 1

The clustering algorithm is sensitive to variations in the numerical ranges of different features. Hence, the obtained clusters can be negatively influenced by the different magnitude of the trace element contents. Therefore, the clustering was

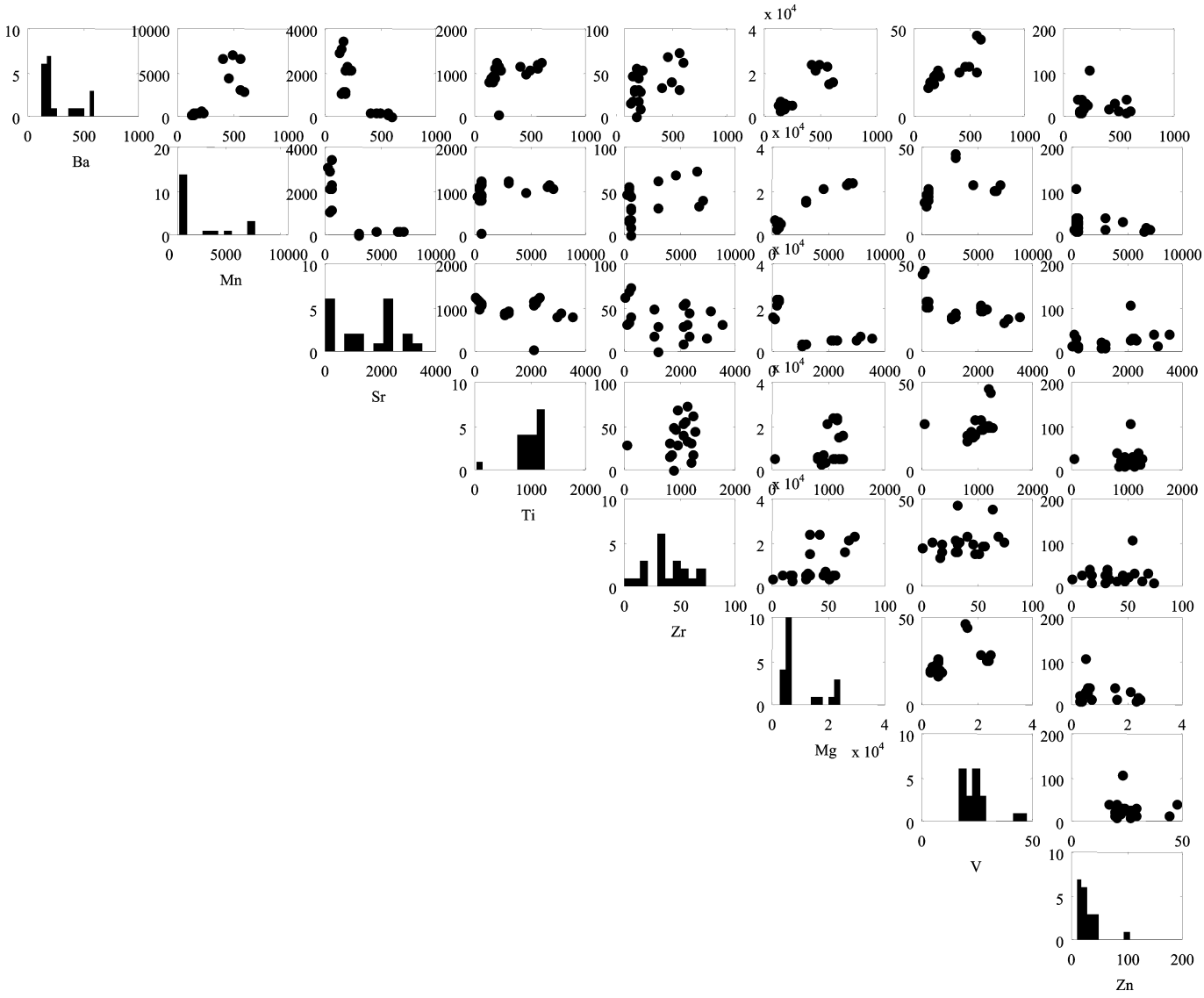

Figure 1 Histogram and covariance of the trace element content of South African clinkers given in mg/kg.

performed based on normalized data, where all transformed features have zero mean and unit variance,

$$\tilde{x}_{j,k} = \frac{x_{j,k} - \bar{x}_j}{\sigma_j}$$

where $\bar{x}_j$ represents the mean, σ_j the variance of the jth feature (trace element).

The dendogram that was constructed based on this data, is shown in Fig. 2. The separation of clinkers produced in different factories is not fully satisfactory, as Factory 4 forms no compact group. This can be explained by the fact that the noisy measurements are analysed in a high (eight) dimensional space.

4.2 Model 2

The second model firstly projects the data into the two-dimensional space of the first principal components and then applies the clustering algorithm. The projected data is shown in Fig. 3. Because of this projection, the structure of the data can also be visually inspected. It can be seen that the clinkers produced in factory SA1 and SA4, and SA2 and SA3 are difficult to separate. The conclusion of this visual inspection is reflected from the analysis of the dendogram depicted in Fig. 4.

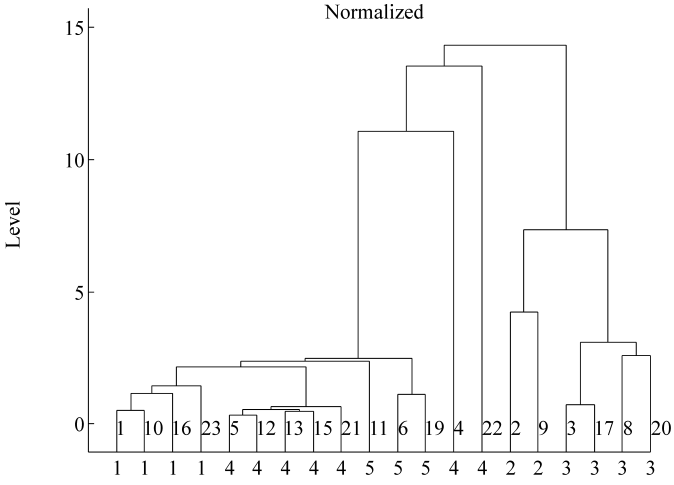

Figure 2 Dendogram of South African clinkers constructed from normalized data. The y-axis shows the distances between clusters. Numbers at the bottom are the factory codes, while numbers within the dendrogram are the sample codes.

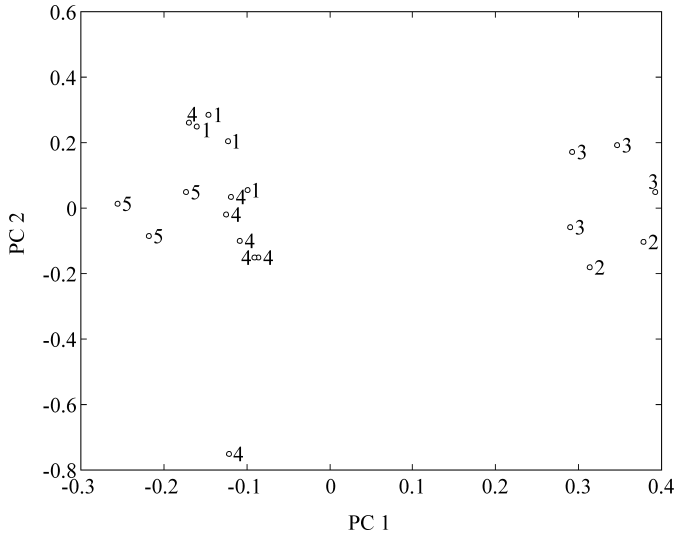

Figure 3 Principal component projection of the data. The numbers denote the code of the manufacturing factory.

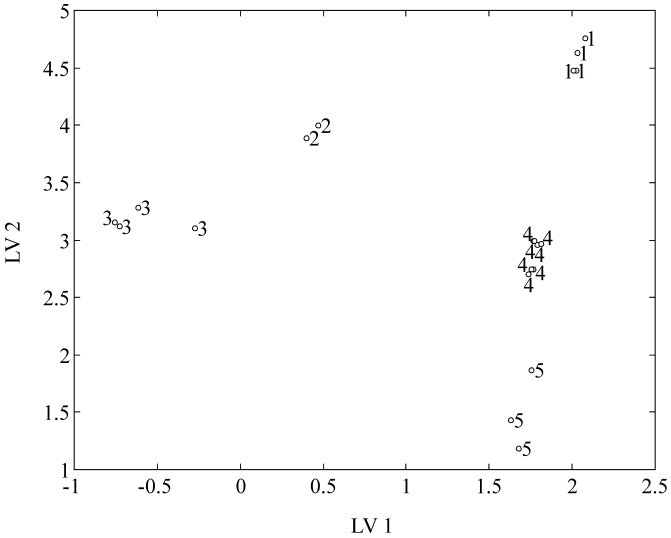

Figure 5 Latent variable projection of the data. The numbers denote the code of the manufacturing factory.

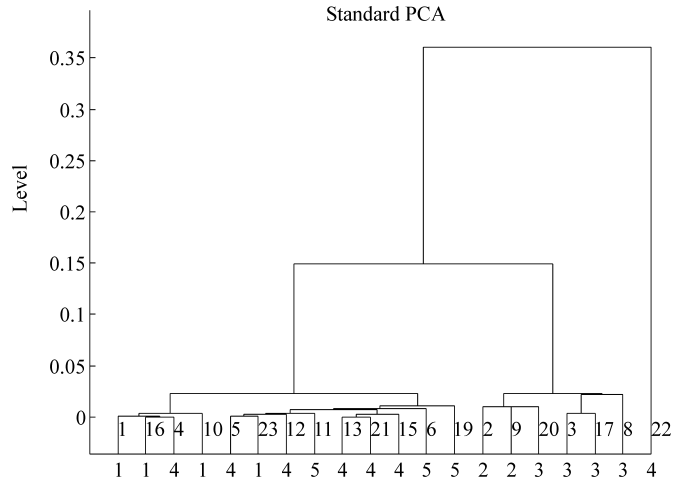

Figure 4 Dendogram of South African clinkers constructed from data projected into the first two principal components. The *y*-axis shows the distances between clusters. Numbers at the bottom are the factory codes, while numbers within the dendrogram are the sample codes.

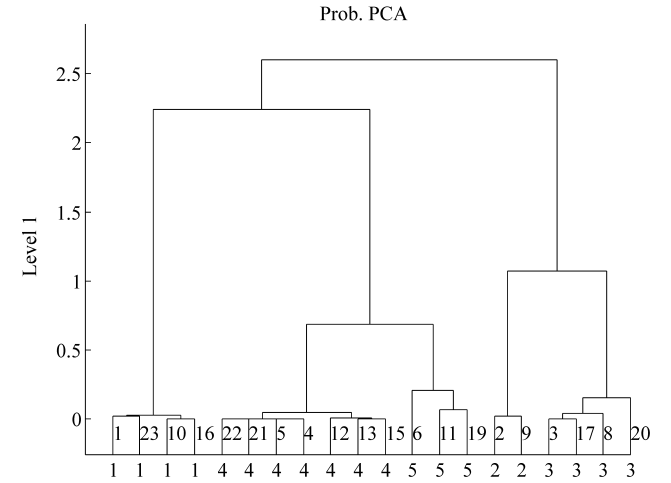

Figure 6 Dendogram of South African clinkers constructed from the two-dimensional latent variables. The *y*-axis shows the distances between clusters. Numbers at the bottom are the factory codes, while numbers within the dendrogram are the sample codes.

4.3. Model 3

The third model projects the data by the presented probabilistic latent variable model. The two-dimensional space of the latent variables is shown in Fig. 5. It can be seen that the data of the objects (trace element content of clinkers) form five clusters related to different factories. Because of this advantageous mapping, the constructed dendogram perfectly clusters the data (Fig. 6).

5. Conclusions

The trace element content of clinkers (and possibly of cements) can be used for the qualitative identification of the manufacturing factory. Hierarchical clustering have been used to achieve this aim. It turned out that normalization of the data is not enough to obtain reliable clustering. To reduce the uncertainty due to the measurement noise and the small number of samples, an advanced projection algorithm has to be used to map the data into a smaller dimensional space. For this purpose standard principal component analysis can be used. However, this method also gave unsatisfactory results, due to the complexity of the PCA compared to the small number of data. The presented probabilistic factor analysis model has the capacity to control the model complexity through the choice of the number of latent variables by limiting the number of parameters used to define the covariance structure of the data. This enables models to be constructed in high-dimensional spaces where fully parameterized covariance matrices would be hopelessly underconstrained. The good performance of this approach emphasizes that the latent variable model has considerable potential for the analysis of trace element content of clinkers and the proposed method is useful to determine the origin of the clinker. A detailed description of the data analysis tools presented in the paper helps with the implementation of the algorithms and a still easier program has been written for this purpose, which can be downloaded (http://www.fmt.vein.hu/softcomp).

Acknowledgements

The financial support of OTKA (Hungarian National Research Foundation), No. T026307 is gratefully acknowledged. Thanks are due to members of the Technical Committee '180-QIC' (Qualitative Identification of Clinkers and Cements) of RILEM (Réunion Internationale des Laboratoires d'Essais et de Recherches sur les Matériaux et les Constructions) for collecting composite average clinker samples. Janos Abonyi is grateful for a

Janos Bolyai Research Fellowship from the Hungarian Academy of Science. The rest of the authors wish to express their thanks to the National Research Foundation of South Africa and the Technikon Pretoria for financial assistance.

References

1 R.L. Goguel and D.A. St John, *Cem. Concr. Res.,* 1993, **23**(1), 59; *Cem. Concr. Res.,*1993, **23**(2), 283.

2 J.C. Miller and J.N. Miller, *Statistics for Analytical Chemistry, Chapter 7.13: Pattern Recognition,* Ellis Horwood, New York, USA 1984.

3 F.D. Tamas, *World Cement,* 1996, **27,** 75

4 F.D. Tamás and É. Kristóf-Makó, Chemical 'fingerprints' in Portland cement clinkers, in *Advances in Building Materials Science – Festschrift Wittmann* (A. Gerdes, ed.), Aedificatio Publishers, Freiburg – Unterengstringen, Germany, 1996, pp. 217–228.

5 F.D. Tamás, A. Tagnit-Hamou and J. Tritthart, Trace elements in clinker and their use as 'fingerprints' to facilitate their qualitative identification, In *Materials Science of Concrete* (M. Cohen, S. Mindess, J. Skalny, eds.) – The Sidney Diamond Symposium, Honolulu, HI, American Ceramic Society, Westerville OH, September 1998, pp. 57–69

6 F.D. Tamás and J. Abonyi, *Cem. & Concr. Res.,* 2002, **32** (8), 1319–1323.

7 B.K. Lavine, Clustering and classification of analytical data. *Encyclopedia of Analytical Chemistry,* John Wiley, New York, 2000.

8 P.D. Wentzell and M.T. Lohnes, *Chemometrics and Intelligent Laboratory Systems,* 1999, **45**, 65.

9 D.T. Andrews and P.D. Wentzell, *Anal. Chim. Acta,* 1997, **350,** 341.

10 M.E. Tipping and C.M. Bishop, *Probabilistic Principal Component Analysis. Technical Report,* NCRG/97/010, 1997.

11 M.E. Tipping and C.M. Bishop, *Mixtures of Probabilistic Principal Component Analysis. Technical Report,* NCRG/98/003, 1998.

12 G. Barkó, J. Abonyi and J. Hlavay, *Anal. Chim. Acta,* 1999, **398** (2–3), 219.