Clasificación de clínkeres basándose en sus elementos traza. Aplicación a clínkeres españoles

Trace elements based classification on clinkers. Application to spanish clinkers

F. D. TAMÁS(*), J. ABONYI(**), F. PUERTAS(***)

(*) Dept. Silicate and Materials Engineering. Univ. de Veszprém, HUNGRIA
(**) Dept. Process Engineering. Univ. de Veszprém, HUNGRIA
(***) Instituto de Ciencias de la Construcción Eduardo Torroja (CSIC), Madrid, ESPAÑA

Fecha de recepción: 10-VII-2001

RESUMEN

En el presente trabajo se describe el procedimiento de identificación cualitativa de clínkeres españoles con el objeto de determinar su origen (fábrica). Esa clasificación de los clínkeres se basa en el contenido de sus elementos traza. Se analizaron 15 clínkeres diferentes procedentes de 11 fábricas de cemento españolas, determinándose los contenidos en Mg, Sr, Ba, Mn, Ti, Zr, Zn y V. Se ha diseñado un sistema experto mediante un árbol de decisión binario basado en los datos recogidos. La clasificación obtenida fue examinada mediante la validación cruzada de 10 valores. Los resultados obtenidos muestran que el modelo propuesto es válido para identificar, de manera fácil, un sistema experto capaz de determinar el origen de un clínker basándose en el contenido de sus elementos traza.

SUMMARY

The qualitative identification to determine the origin (i.e. manufacturing factory) of Spanish clinkers is described. The classification of clinkers produced in different factories can be based on their trace element content. Approximately fifteen clinker sorts are analysed, collected from 11 spanish cement factories to determine their Mg, Sr, Ba, Mn, Ti, Zr, Zn and V content. An expert system formulated by a binary decision tree is designed based on the collected data. The performance of the obtained classifier was measured by ten-fold cross validation. The results show that the proposed method is useful to identify an easy-to-use expert system that is able to determine the origin of the clinker based on its trace element content.

1. INTRODUCCIÓN

El contenido de los elementos traza en los clínkeres es de gran interés científico y puede ser utilizado para resolver problemas prácticos, como, por ejemplo, determinar el origen del clínker. El primer trabajo en un tema similar fue publicado en 1993 por Goguel and StJohn (1), quienes determinaron las concentraciones de Ba, Sr y Mn de cementos portland en hormigones de Nueva Zelanda. Este primer acercamiento sugirió la necesidad de emplear métodos estadísticos avanzados, denominados "pattern recognition" or "fingerprinting", para facilitar la identificación cualitativa (2).

1. INTRODUCTION

The trace element content of clinkers is of high 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 of similar topics was published in 1993 by Goguel and StJohn (1), showing the Ba, Sr and Mn concentration of portland cements in New Zealand concretes. This first attempt suggests that advanced statistical methods, called "pattern recognition" or "fingerprinting" can help qualitative identification (2).

No obstante, la identificación cualitativa requiere bases de datos que permitan comparar los contenidos de los elementos traza de un clínker o cemento desconocido con aquellos de muestras ya conocidas. Se puede consultar información adicional de los contenidos de elementos traza en clínker y cementos en publicaciones anteriores (3, 4, 5). En estos trabajos se ha demostrado que no todos los elementos traza pueden ser utilizados como "fingerprinting"; la selección debe seguir ciertos principios. Un aspecto importante en la selección a es que los elementos traza de "valor dactilogramático" deben proceder de las principales materias primas (calizas, margas, arcillas) y no del combustible, ni de la línea del horno ni de los revestimientos de los molinos. Más recientemente se han utilizado 6 elementos en la caracterización de los clínkeres: los ya indicados por Goguel and St. John (1) más Mg, Ti y Zr (4, 5). El Zn y el V no tienen valor dactilogramático (pueden proceder de los combustibles si se utilizan neumáticos u otros combustibles), pero su cantidad puede ser interesante a la hora de valorar el comportamiento del cemento.

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 too (3, 4, 5). In these papers it was shown, that not all trace elements can be used for fingerprinting; selection must follow certain principles. The most important item of selection: trace elements of "dactylogrammatic value" should come from the main raw material (limestone, marl, clay) and not from the fuel, from furnace lining or from grinding media wear, and some other principles should be observed as well. More recently 6 elements were used to characterize clinkers: besides those used by Goguel and St. John (1) the Mg, Ti and Zr contents were measured too (4, 5). Zn and V have no dactylogrammatic value (they come from the fuel, if waste tyres or special sorts of heavy fuel oil are used, resp.), but their quantity can be interesting in cement performance.

En trabajos previos (6) el valor dactilogramático de los elementos traza ya fue descrito junto con información detallada sobre la preparación de las muestras y los procesos de análisis. Se determinaron las desviaciones medias y estándar de los ocho elementos traza considerados (Mg, Sr, Ba, Mn, Ti, Zr, Zn y V). A partir del análisis de más de 200 muestras se ha calculado el contenido "estándar" de los elementos traza y, con el fin de facilitar la visualización de dichos contenidos, se presentó un método gráfico ("Star Plotting"), que permite comparar cualquier clínker con el estándar propuesto.

In our previous paper (6) the dactylogrammatic value of trace elements was described, jointly with detailed data on sample preparation and analysis; averages and standard deviations of eight trace elements (Mg, Sr, Ba, Mn, Ti, Zr, Zn and V) were tabulated. Based on >200 samples, a "standard" trace element content was calculated and, in order to facilitate the visualisation of the trace element content, a graphical method ("Star Plotting") was presented, where every clinker is compared to the proposed standard.

Se han aplicado diferentes métodos estadísticos avanzados y de patrones de reconocimiento para agrupar los clínkeres, de manera que los datos analíticos sean transformados en componentes básicos y se puedan construir dendogramas (3, 4, 5).

Among the wide range of advanced statistical and "pattern recognition" methods, hierarchical clustering technique have been applied for the clustering of clinkers, where the analytical data were transformed by principal component analysis and dendograms were constructed for cluster formation (3, 4, 5).

Recientemente, se ha propuesto (7) una nueva aproximación para identificar e interpretar la regla basada en sistemas expertos. Las reglas basadas en sistemas expertos se aplican con frecuencia para la clasificación de problemas relacionados con la detección de fallos, en biología, medicina, etc. La lógica difusa ("fuzzy logic") mejora la clasificación y la decisión de los sistemas permitiendo el uso de clases superpuestas y mejorando la interpretación de los resultados, proporcionando más claridad en el clasificador y en el proceso de identificación (8). En este trabajo asumimos que, durante el proceso de identificación de las muestras, las fábricas no eran

In (7) a new approach has been proposed to identify an easily interpretable rule-based expert system. Rule-based expert systems are often applied to classification problems in fault detection, biology, medicine, etc. Fuzzy logic improves classification and decision support systems by allowing the use of overlapping class definitions and improves the interpretability of the results by providing more insight into the classifier structure and decision making process (8). In this paper, we assumed that during the training of the classifier the factories of the samples were not known. Hence, for the

conocidas. De aquí, que para la identificación de la regla basada en sistemas expertos se haya aplicado la segmentación por grupos ("fuzzy clustering") de manera aleatoria, que está entre los métodos que no utilizan una clase identificadora; en (9) puede verse un ejemplo de segmentación por grupos.

En el presente trabajo se ha seguido una aproximación diferente. La regla, basada en el sistema experto, se representa mediante el desdoblamiento del árbol de decisión binaria que está identificada mediante el algoritmo C4.5 (10, 11).

El resto del artículo se ha estructurado como sigue: en el siguiente apartado se describe el trabajo experimental de los análisis químicos de los clínkeres españoles; en el tercer apartado se describen brevemente los fundamentos del árbol de decisión binaría y el algoritmo de inducción C4.5; en el apartado 4, previo a las conclusiones, se presentan los resultados de la aplicación de esta metodología a la identificación cualitativa de los clínkeres españoles. Este ejemplo demuestra que el modelo propuesto es útil y fácil de aplicar para determinar el origen de un clínker.

2. EXPERIMENTAL

2.1. Materiales

Para la identificación ("fingerprinting") cualitativa de los clínkeres, obviamente se necesita partir de muestras de clínker muy bien definidas. Para obtener una base de datos realmente valiosa se creó en 1996 y auspiciado por la RILEM (Réunion Internationale des Laboratoires d'Essais et de Recherches sur les Matériaux et les Constructions) un Comité Técnico, denominado "QIC" (Identificación Cualitativa de Clínkeres y Cementos) (TC 180/QIC). Este proyecto ha permitido coleccionar una amplia gama de muestras procedentes de 8 países (Austria. Portugal, Sudáfrica, Eslovaquia, Eslovenia, España, Suiza y Reino Unido). Se analizaron más de 200 muestras(*). En este artículo se presentan únicamente los datos referentes a los clínkeres españoles. 15 clínkeres distintos fueron analizados (procedentes de 11 fábricas de cemento españolas), determinándose su contenido en Mg, Sr, Ba, Mn, Ti, Zr, Zn,y V. La metodología analítica empleada se describe a continuación.

(*) Se pueden obtener datos analíticos detallados solicitándoselos al autor, pero no se da información sobre la compañía, fábrica o muestreo, únicamente se informa del país y del código empleado.

identification of the rule based expert system fuzzy clustering has been applied that is among unsupervised learning methods, since it does not use a priori class identifiers. A chemometric example for the effective use of fuzzy clustering can be found in (9).

In this paper a different approach have been followed. The rule-based expert system is represented by a crisp binary decision tree that is identified by the supervised C4.5 learning algorithm (10, 11).

The rest of this paper is organised as follows: the next section deals with the experimental details of the chemical analysis of spanish clinkers: in the third section the concept of binary decision tree is presented and the C.4.5 decision tree induction algorithm is briefly described; In section 4, before the conclusions, the results of the factual application example of the qualitative identification spanish clinkers is given. This example illustrates the proposed method is useful and easily applicable 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 are necessary. To obtain such informative database, a Technical Committee "QIC" (Qualitative Identification of Clinkers and Cements) has been established in 1996, under the auspices of RILEM (Réunion Internationale des Laboratoires d'Essais et de Recherches sur les Matériaux et les Constructions) (TC 180/QIC). This project enabled to collect composite average samples from 8 further countries (Austria, Portugal, South Africa, Slovakia, Slovenia, Spain, Switzerland and United Kingdom). Over 200 samples had arrived to date, and analysed(*). This paper focuses only on spanish clinkers. Approximately fifteen clinker sorts have been analysed collected from eleven spanish cement factories to determine their Mg, Sr, Ba, Mn, Ti, Zr, Zn and V content. The details of the chemical analysis is given bellow.

(*) *Detailed analytical data can be obtained from the author. By the request of companies, factory and sampling data cannot be revealed, only the country and a code number.*

2.2. Métodos

La cantidad de muestra de clínker recibida en los laboratorios de análisis fue de 2-3 kg aproximadamente, en forma de nódulos. Estas muestras fueron machacadas y molidas, en molinos de centrífuga, de acuerdo a métodos estándar de muestreo. El tamaño final de las partículas (inferior a 63 μm) se consiguió moliendo la muestra a mano en mortero de ágata. En estudios previos con cuarzo puro se había demostrado que la abrasión debida al método de molienda o al molino empleado no contaminaba la muestra en los elementos a analizar. Aproximadamente 1 g de muestra, pesada con precisión, se disolvía en ácido clorhídrico, se precipitaba el SiO_2 que se filtraba y lavaba. El filtrado era posteriormente analizado por ICP-ES (Espectrometría de Emisión con Fuente de Plasma de Acoplamiento Inductivo). El ICP-ES es una técnica muy válida cuando se precisan valores de ppm absolutos, ya que la calibración se puede hacer a partir de reactivos químicos de concentración conocida. Todos los análisis de los clínkeres se realizaron por duplicado; cuando la diferencia entre los valores era superior al 10%, se repetía la preparación y el análisis.

Se utilizaron dos equipos ICP-ES. Al inicio el equipo era un ARL/3410 y posteriormente un GBC-Integra-XM tipo ICP, con antorcha de miniplasma y generadores de radiofrecuencia, 27 /(40) MHz, de 650 W(2 kW) power (los datos entre paréntesis corresponden al equipo nuevo). El rango espectral iba de 165 a 800 nm. Sistema de computacion EPIC (Evolutionary Program for Instrument Control) software IBM PS/2 con un PC DOS 3.0. Las longitudes de onda usadas en este estudio en nm fueron: Mn=257.610, Mg=279.553, Sr=407.771, Ba=455.403, Ti=336.121, Zr=349.621, Zn=213.856, V=310.230.

2.3. Obtención de la Base de Datos

En la Tabla 1 se muestran los contenidos de los elementos traza analizados en los clínkeres españoles.

En la mayoría de los casos se analizaron 4 clínkeres diferentes procedentes de cada fábrica, a excepción de la fábrica 1 de la que se analizaron 11 clínkeres, de la fábrica 4 con 8 clínkeres y de la fábrica 8 con 2 clínkeres. En la Tabla 2 se presentan los valores medios y las desviaciones estándar calculadas a partir de los datos analíticos obtenidos para los 8 elementos considerados en las 53 muestras analizadas.

2.2. Methods

The mass of clinker samples, which arrived to this laboratory was aprrox. 2-3 kg, usually as noncrushed nodules. This was crushed and a smaller average taken, according to sampling standards, and ground in a centrifugal mill. The final size reduction, (to pass sieve 63 μm) was handmade, in an agate mortar. (A previous experiment, with pure quartz showed that the abrasion of grinding media or mill lining did not bring any considerable pollution into the sample of the elements analysed). Approx. 1 g of exactly weighed sample was dissolved in r.g. hydrochloric acid, SiO_2 precipitated, filtered, washed and the filtrate analysed by ICP-ES (Inductively Coupled Plasma Emission Spectrography). ICP-ES is advantageous when absolute ppm values are needed, as standardization can be done by reagent grade chemicals of known concentration. Duplicate samples were prepared of all clinkers for analysis, and if the difference was > 10%, sample preparation and analysis was repeated.

Details of ICP-ES analysis: in the beginning period an ARL/ 3410, later a GBC-Integra-XM type ICP spectrometer was used, with mini-plasma torch, using radiofrequency generators, 27 /40/ MHz, of 650 W /2 kW/ power. /Data in parentheses refer to the new apparatus/. Spectral range: 165-800 nm. Computation: EPIC (Evolutionary Program for Instrument Control) software, IBM PS/2 computer, with a PC DOS 3.0 system. The wavelengths used in this study (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.

2.3. Obtained database

The measured trace element content of the Spanish clinkers are shown in Table 1.

Mostly four different clinkers from each factory have been analyzed. Only factory 1, 4. and 8 were exception with sample number 11, 8 and 2, respectively. Averages and standard deviations (std) calculated using the obtained data of 53 samples for the 8 considered elements are shown in Table 2.

TABLA 1 / *TABLE 1*

Contenido de elementos trazas en los clínkeres españoles (mg/kg)

Trace element content of spanish clinkers (mg/kg)

Código *Code*	Ba	Mn	Sr	Ti	Zr	Mg	V	Zn	Fábrica *Factory*
1	84	26	206	621	23	2278	28	17	*1*
2	143	15	317	235	15	2191	47	11	*4*
3	98	197	195	1235	49	5004	132	105	*6*
4	339	435	242	1336	49	7204	225	40	*10*
5	288	185	329	1242	44	5982	166	43	*9*
6	440	147	588	1119	66	12584	106	184	*5*
7	112	202	269	1115	56	6253	75	243	*4*
8	117	89	1179	1023	39	8910	181	26	*2*
9	285	268	249	1280	71	18417	176	50	*3*
10	283	273	321	591	36	4005	204	296	*8*
11	309	144	413	1169	47	7186	177	111	*11*
12	118	162	281	1246	66	6900	170	245	*7*
13	235	338	352	1206	67	9641	215	22	*1*
14	192	160	345	1170	67	9745	216	22	*1*
15	101	171	210	1125	50	4401	140	149	*6*
16	195	1024	302	824	57	5913	47	26	*1*
17	235	333	295	1402	69	21115	182	21	*3*
18	293	184	449	1225	43	7291	166	192	*11*
19	67	16	312	175	12	1750	53	11	*4*
20	108	88	1253	1024	37	8840	247	35	*2*
21	442	185	301	1335	48	6114	155	82	*9*
22	67	152	293	1041	38	6219	145	559	*7*
23	357	606	370	1203	63	7239	137	64	*10*
24	182	208	445	1177	67	9233	143	45	*1*
25	181	247	318	761	13	4284	140	404	*8*
26	266	131	640	1211	65	14826	121	329	*5*
27	47	27	214	723	36	2202	309	11	*1*
28	121	213	288	1379	38	4826	87	304	*4*
30	275	125	282	1114	59	6213	198	59	*1*
31	257	1430	352	708	41	4333	214	12	*1*
32	57	18	287	446	25	1631	51	36	*4*
33	98	136	162	1219	49	3639	161	128	*6*
34	155	136	471	1125	58	11204	92	436	*7*
35	208	398	341	1424	75	22385	202	23	*3*
36	160	132	484	1130	60	11343	95	412	*2*
37	107	106	1389	1137	53	10861	297	21	*5*
38	99	211	315	1333	52	5882	95	387	*4*
39	254	171	405	1404	49	6721	197	34	*9*
40	334	662	320	1238	69	8160	186	92	*10*
41	187	131	460	1151	53	9078	173	45	*11*
42	239	186	516	1326	46	9157	194	182	*1*
43	228	1152	273	700	56	6615	161	37	*1*
44	80	29	189	630	32	2274	288	26	*1*
45	94	76	1243	1023	36	10204	234	40	*2*
46	278	581	332	1519	46	22061	196	21	*3*
47	90	165	239	1153	44	8229	82	327	*4*
48	57	14	298	305	15	1751	51	8	*4*
49	214	113	705	1107	66	13769	75	340	*5*
50	82	134	179	1036	39	3740	105	85	*6*
51	101	167	272	1267	53	10027	113	444	*7*
52	251	169	411	1378	46	6614	195	40	*9*
53	331	637	312	1190	61	7834	176	71	*10*
54	232	174	493	1275	46	8702	188	188	*11*

TABLA 2 / *TABLE 2*

Valores medios y desviaciones estándar de los clínkeres investigados agrupados por fábricas (mg/kg)

Average values and standard deviations of investigated clinkers grouped by factories (mg/kg)

Fábrica / Factory		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
# data		11	4	4	8	4	4	4	2	4	4	4
Ba	Media / Mean	178.4	106.5	251.5	93.3	270.0	94.8	110.3	232.0	308.8	340.3	268.3
	Std	75.9	9.5	36.5	31.5	121.3	8.6	36.6	72.1	90.4	11.6	38.5
Mn	Media / Mean	422.7	89.8	395.0	106.8	130.8	159.5	154.3	260.0	177.5	585.0	172.0
	Std	517.0	12.3	134.9	98.4	13.9	30.2	13.7	18.4	8.7	102.6	19.4
Sr	Media / Mean	310.9	1266.0	304.3	290.6	604.3	186.5	329.3	319.5	361.5	311.0	467.8
	Std	90.8	88.3	41.9	26.6	93.4	20.7	94.9	2.1	55.0	52.7	45.9
Ti	Media / Mean	911.3	1051.8	1406.3	767.6	1141.8	1153.8	1169.8	676.0	1339.8	1241.8	1248.8
	Std	248.1	56.8	98.3	523.0	47.1	92.3	106.2	120.2	71.1	66.0	67.3
Zr	Media / Mean	50.7	41.3	65.3	32.1	64.3	46.8	53.8	24.5	46.8	60.5	45.5
	Std	15.4	7.9	13.1	17.7	2.9	5.2	11.8	16.3	2.2	8.4	1.7
Mg	Media / Mean	6138.6	9703.8	20994.5	4064.1	13130.5	4196.0	8587.5	4144.5	6357.8	7609.3	8084.0
	Std	3039.2	994.2	1800.8	2567.8	1502.9	635.9	2406.6	197.3	364.3	467.3	994.7
V	Media / Mean	181.1	239.8	189.0	67.6	99.3	134.5	130.0	172.0	178.3	181.0	181.3
	Std	86.5	47.7	12.1	19.2	19.4	23.2	34.4	45.3	21.0	36.2	12.4
Zn	Media / Mean	29.3	30.5	28.8	165.9	316.3	116.8	421.0	350.0	49.8	66.8	168.3
	Std	15.3	8.6	14.2	164.6	95.5	27.8	130.1	76.4	21.8	21.4	38.4

3. CONSTRUCCIÓN DE LOS ÁRBOLES DE DECISIÓN

3.1. Aplicación de los Árboles de Decisión para la clasificación

La adquisición de conocimientos a partir de los ejemplos, es decir, la adquisición de conceptos, es una de las ramas más importantes de la técnica de aprendizaje, que, en general, es considerada como el cuello de botella en el desarrollo de los sistemas expertos. Para este fin se ha desarrollado una amplia gama de modelos y algoritmos de identificación. En el

3. CONSTRUCTION OF DECISION TREES

3.1. Application of decision trees for classification

Learning from examples, i.e. concepts acquisition, is one of the most important branches of machine learning that has been generally regarded as the bottle-neck of expert system development. For this purpose a wide range of models and identification algorithms have been developed. Among them,

presente trabajo se han aplicados los árboles de decisión para crear una regla base del clasificador.

El árbol de decisión binaria comprende dos tipos de nodos: (i) nodos internos que tienen dos ramas, y (ii) nodos terminales sin ramas. Cada nodo interno está asociado con una función de decisión para indicar que nodo siguiente debe ir. Cada nodo terminal representa la salida de una entrada dada que lleva a este nodo, es decir, en los problemas de clasificación cada nodo terminal contiene la etiqueta de la clase indicada. En la Figura 1 se da un ejemplo ilustrativo para un árbol de decisión, donde el árbol de decisión define el problema de clasificación de dos clases basado en dos variables de entrada, x_1 y x_2. La Figura 2 representa un problema espacial, que pone de manifiesto, que el árbol de decisión puede representarse mediante reglas del tipo:

Si (x_1 es mayor que 2) y (x_1 es menor que 5) y (x_2 es menor que 5) entonces Clase 2

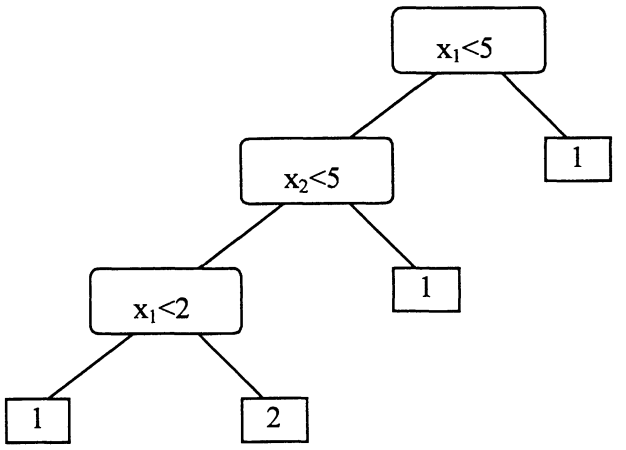

Figura 1.- Ejemplo ilustrativo de un árbol de decisión binaria.

Figure 1.- Illustrative example for a binary decision tree.

through the paper binary decision trees are applied to create rule-based of the classifier.

A binary decision tree consists of two type of nodes: (i) internal nodes having two branches, and (ii) terminal nodes without branches. Each internal node is associated with a decision function to indicate which node to visit next. Each terminal node represents the output of a given input that leads to this node, i.e. in classification problems each terminal node contains the label of the predicted class. An illustrative example for a decision tree is given in Figure 1, where the decision tree defines a two-class classification problem based on two input variables, x_1 and x_2. As the problem space illustrated by Figure 2. shows, the decision tree can be also represented by rules like,

If (x_1 is more than 2) and (x_1 is less than 5) and (x_2 is less than 5) then Class 2

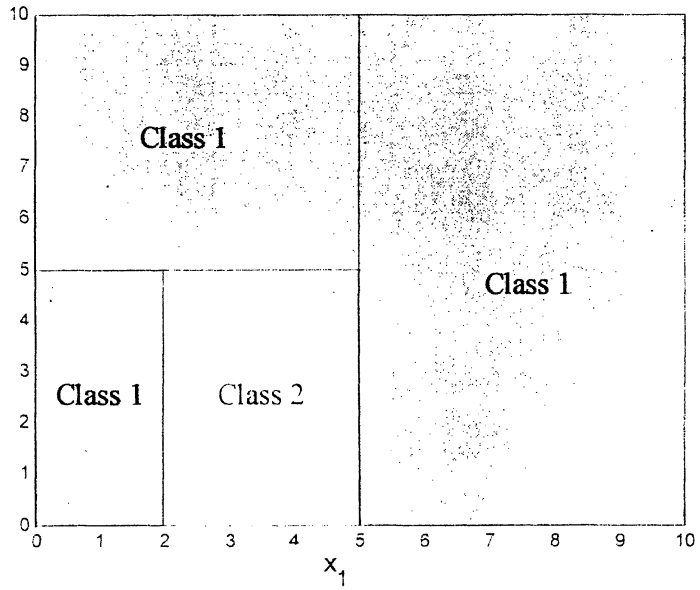

Figura 2.- Ilustración de un árbol de decisión de un problema de clasificación espacial de dos clases.

Figure 2.- Two-class classification problem space of the illustrative decision tree.

3.2. Algoritmo C4.5

Los algoritmos para la construcción del árbol de decisión dan lugar a árboles de decisión a partir de un conjunto D de casos. Estos algoritmos realizan una partición del conjunto de los datos D en subgrupos D_1, D_2, ..., D_M mediante una serie de ensayos T con valores recíprocos T_1, T_2, ..., T_M; donde D_i contiene aquellos casos que tienen valores de ensayos T_i. El C4.5 es un árbol de decisión binaria que genera un algoritmo (10) que se aplica como sigue:

Para datos numéricos (continuos) el ensayo se escribe como $x_j < t$. Los valores umbrales de t vienen

3.2. C4.5 algorithm

The decision tree construction algorithms generate decision trees from a set D of cases. Theses algorithms partition the D data set into subsets D_1, D_2, ..., D_M by a set of tests T with mutually outcomes T_1, T_2, ..., T_M where D_i contains those cases that have outcome T_i. The C4.5 is a binary decision tree generating algorithm [10] and applied in the following.

For numeric (continuous) attributes the attribute test is written as $x_j < t$. The t-thresholds are selected

seleccionados de acuerdo a un criterio de partición. El criterio de partición por defecto utilizado por C4.5 es el criterio de ganancia, que se define como una medida de la información que tiene en cuenta las diferentes probabilidades de los resultados. La razón de ganancia se explica como la incertidumbre residual a cerca de la clase a la cual pertenece un caso en D:

based on a splitting criterion. The default splitting criterion used by C4.5 is the gain ratio, as an information-based measure that takes into account different probabilities of the outcomes. The gain ratio is explained as follows. The residual uncertainty about the class to which a case in D belongs can be expressed as:

$$\text{Infor (D)} = -\sum_{j=1}^{M} p(D,j)\log_2(p(D,j))$$

Donde p (D, j) indica la proporción de las clases que pertenecen a la clase j.

where p(D,j) denotes the proportion of classes in D that belong to the jth class.

La ganancia de información mediante un ensayo viene fuertemente afectada por el número de valores y es máxima cuando hay una clase en cada subgrupo D_i:

The information gained by a test is strongly effected by the number of outcomes and is maximal when there is one class in each subset D_i:

$$\text{Ganancia (D, T)} / \textit{Gain (D, T)} = \textit{Infor (D)} - \sum_{i=1}^{M} \frac{|D_i|}{|D|} \textit{Infor (D}_i\textit{)}$$

Donde $|D_i|$ indica el cardinal del conjunto de datos D_i.

where $|D_i|$ denotes the cardinality of the D_i data set.

Por otro lado, la información potencial obtenida mediante la partición de un conjunto de casos está basado en el conocimiento del subgrupo D_i, en el cual un caso es eliminado:

On the other hand, the potential information obtained by partitioning a set of cases is based on knowing the subset D_i, into which a case falls:

$$\text{Partición (D, T)} / \textit{Split (D, T)} = -\sum_{i=1}^{M} \frac{|D_i|}{|D|}\log_2\left(\frac{|D_i|}{|D|}\right)$$

La información de partición tiende a aumentar con el número de valores del ensayo. El criterio de la relación de la ganancia evalúa la conveniencia de un ensayo por medio del cociente entre la información de la ganancia y la información de la partición. La relación de la ganancia se determina para cada ensayo y, entre aquéllas con ganancia media mínima, se selecciona la partición con relación de ganancia de valor máximo (10).

This splet information is tends to increase with the number of outcomes of a test. The gain ratio criterion assesses the desirability of a test as the ratio of its information gain to its split information. The gain ratio of every possible test is determined and, among those with at least average gain, the split with maximum gain ratio is selected (10).

Los resultados obtenidos de la partición permiten construir árboles representativos con los datos del ensayo. En las aplicaciones prácticas, con frecuencia los datos contienen ruido (*noise*), lo cual, en general, conduce a árboles demasiado complejos. De aquí, que los métodos de construcción de un árbol con mayor decisión reduzcan el árbol inicial mediante subárboles de identificación, que tienen poca repercusión en la precisión.

The recursive partition strategy results in trees that are consistent with the training data. In practical applications, data contains often noise, which leads generally to too complex trees. Hence, most decision tree construction methods prune the initial tree by identifying sub-trees that contribute only a little to the predictive accuracy by replacing these by a leaf.

4. APLICACIÓN EN LA IDENTIFICACIÓN DE LOS CLÍNKERES

4. APPLICATION FOR IDENTIFICATION OF CLINKERS

Se ha aplicado en el método propuesto el programa MATLAB© que es para uso científico. El programa

The proposed method has been implemented in MATLAB© that is a scientific computing programming

puede cargarse desde la página del autor J. Abonyi: www.fmt.vein.hu/softcomp.

La precisión del modelo se mide en términos del número de desclasificados (mal clasificados). El comportamiento del clasificador fue medido mediante la validación cruzada de diez turnos. Esto significa que los datos se dividen en 10 subgrupos de casos que tienen un tamaño similar y distribución de clase. De cada subgrupo se elimina uno, mientras que los nueve restantes son utilizados para la construcción del clasificador que es posteriormente validado por casos no utilizados en la serie que se ha rechazado. Este método es muy utilizado en la técnica del aprendizaje, donde el número de datos es relativamente pequeño, al igual que ocurre en nuestro estudio.

Cuando todos los elementos traza son utilizados para la clasificación, el árbol de decisión obtenido por C4.5 da una desclasificación de 1,88% (1 mala clasificación) sobre una validación de 10 grupos (98,11% clasificación correcta) con 13 nodos terminales. El clasificador obtenido no utiliza los datos relativos a los contenidos de V, lo que apoya la suposición inicial que el Zn y el V no tienen validez dactilogramática (proceden principalmente del combustible) y no pueden ser utilizados para la identificación de los clínkeres.

El árbol de decisión obtenido basado únicamente en los contenidos de Ba, Mn, Sr, Ti, Zr y Mg se muestra en la Figura 3. Este árbol de decisión da sólo dos desclasificados, con 11 y 13 nodos terminales. Las excepciones son (fábrica real/fábrica determinada por el clasificador) # 31 and # 39.

Analizando los resultados puede deducirse que el sistema experto utilizado es innecesariamente complejo, y que no todos los elementos seleccionados pueden ser necesarios para identificar el clínker. A partir de los análisis del árbol de la Figura 3, se seleccionaron tres elementos (Ba, Sr y Mg) para experimentos posteriores. Los datos en tres dimensiones de los resultados obtenidos se muestran en la Figura 4. Como se desprende de esa figura, el problema de clasificación de los clínkeres no es trivial, especialmente, en la identificación de los clínkeres de las fábricas 1, 5 y 10.

A diferencia a cuando se utilizan pocos elementos traza, el clasificador resultante da sólo desclasificaciones # 5 (fábrica 1) y # 26 (fábrica 10) con 14 nodos terminales.

El clasificador obtenido es fácil de usar e interpretar para técnicos e investigadores, aunque no estén familiarizados con el concepto de la técnica de

language. The program can be downloaded from the homepage of author J. Abonyi : www.fmt.vein.hu softcomp.

The model accuracy is measured in terms of the number of misclassifications. The performance of the classifier was measured by ten-fold cross validation. This means that the data is divided into ten sub-sets of cases that have similar size and class distributions. Each sub-set is left out once, while the other nine remaining are applied for the construction of the classifier which is subsequently validated for unseen cases in the left-out sub-set. This method is widely used in machine learning, where the number of training data is relatively small similarly to our case.

When all trace elements are used for classification, the decision tree obtained by C4.5 gives misclassification of 1.88% (1 misclassification) on 10-fold cross validation (98.11% correct classification) with 13 terminal nodes. The resulted classifier does not utilise information about the V content which supports our initial assumption that the Zn and V do not have dactylogrammatic value (they mainly come from the fuel) and they cannot be used for the identification of clinkers.

The decision tree identified based on only Ba, Mn, Sr, Ti, Zr, and Mg is shown in Figure 3. This decision tree gives only two misclassifications, with 11 internal and 13 terminal nodes. The exceptions (real factory/factory determined by the classifier),are and # 31 and #39.

It can happen that the obtained rule-based expert system is unnecessarily too complex as not all the features are needed to identify the clinker. Based on the analysis of the tree shown in Figure 3 three features (Ba Sr, and Mg) were selected for further experiments. The data in this three-dimensional space is depicted in Figure 4. As can be seen, the studied classification problem is not trivial, especially the identification of clinkers produced in factory 1., 5.., and 10 are difficult.

Contrary to the small number of utilised trace elements the resulted classifier gives only to misclassifications (#5 (factory 1) and #26 (factory 10)) with 14 terminal nodes.

The obtained classifier is easy to use and interpret for engineers and researchers, even when they are not familiar with the concept of machine learning.

aprendizaje. Por ejemplo, cada nodo terminal del árbol representa una regla, es decir la primera regla del sistema experto definido por el árbol de decisión dado en la Figura 3, puede ser formulado como:

Si Ba <= 155 y Sr <=214 y Zr <= 37, entonces es la fábrica 1

For instance, each terminal node of the tree can be represented by a rule, e.g. the first rule of the expert system defined by the decision tree given in Figure 3 can be formulated as

If Ba <= 155 and Sr <= 214 and Zr <= 37, then factory 1

```
Ba <= 155 :                          Ba > 155 :
|   Sr <= 214 :                      |   Mg <= 10204 :
|   |   Zr <= 37 : 1                 |   |   Zr <= 48 :
|   |   Zr > 37 : 6                  |   |   |   Zr <= 41 : 8
|   Sr > 214 :                       |   |   |   Zr > 41 :
|   |   Sr > 705 : 2                 |   |   |   |   Mg <= 6900 : 9
|   |   Sr <= 705 :                  |   |   |   |   Mg > 6900 : 11
|   |   |   Mg <= 5982 : 4           |   |   Zr > 48 :
|   |   |   Mg > 5982 :              |   |   |   Ba <= 293 : 1
|   |   |   |   Sr <= 269 : 4        |   |   |   Ba > 293 : 10
|   |   |   |   Sr > 269 : 7         |   Mg > 10204 :
                                     |   |   Mn <= 202 : 5
                                     |   |   Mn > 202 : 3
```

Figura 3.- Árbol de decisión deducido mediante el algoritmo C4.5 para la identificación cualitativa de clínkeres españoles (árbol de decisión II en el texto).

Figure 3.- Decision tree learned by the C4.5 algorithm for the qualitative identification of Spanish clinkers (decision tree II in the text).

5. CONCLUSIONES

El contenido de elementos traza en los clínkeres (y posiblemente en cementos) puede ser usado para la identificación cualitativa de los mismos (es decir, reconocer su fábrica de producción). Para esta finalidad, se han analizados clinkeres españoles para determinar su contenido en Ba, Mn, Sr, Ti, Zr, Zn y V. Los primeros seis elementos proceden de las materias primas y tienen valor dactilogramático, mientras que los dos últimos proceden del combustible (neumáticos usados, aceites pesados, etc.) y no pueden ser utilizados para la identificación.

Para la identificación cualitativa de los clinkeres se designó una regla del clasificador base ensayado mediante el algoritmo C4.5. El estudio realizado ha demostrado que el Ba, Sr y Mg son los elementos traza más relevantes para la identificación "fingerprinting" de los clinkeres.

5. CONCLUSIONS

The trace element content of clinkers (and possibly of cements) can be used for the qualitative identification (i.e. manufacturing factory). For this purpose, several samples from Spain have been analysed to determine their Ba, Mn, Sr, Mg; Ti, Zr; Zn and V content. The first 6 elements come from the main raw materials and are of dactylogrammatic value, while the last two elements mainly come from the fuel (used tires, heavy fuel oil, etc.) and cannot be used for identification.

For the qualitative identification of clinkers a rule base classifier was designed trained by C4.5 algorithm. It has turned out that Ba, Sr and Mg are the most relevant trace elements for fingerprinting of clinkers.

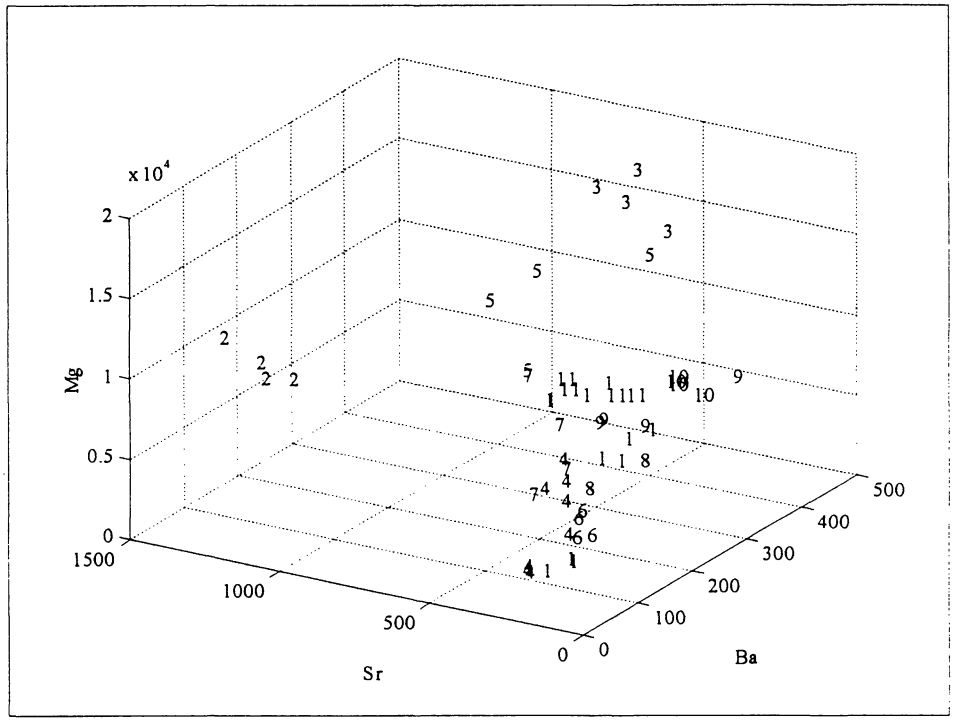

Figura 4.- Contenido de Ba, Sr, y Mg en los clínkeres españoles (53 muestras de 11 fábricas) (mg/kg).

Figure 4.- Ba, Sr and Mg content of Spanish clinkers (53 samples from 11 factories) (mg/kg).

Los resultados han demostrado que el método propuesto es útil para identificar clasificadores compactos que permiten determinar el origen de los clínkeres. Una descripción detallada ayuda en la aplicación del algoritmo de inducción del árbol de decisión; se ha construido un programa de fácil aplicación para este fin y puede ser cargado en (www.fmt.vein.hu/softcomp).

6. AGRADECIMIENTOS

Se agradece sinceramente la ayuda financiera de la OTKA (Fundación de Investigación Nacional Húngara), No.T026307.Se agradece a las fábricas de cemento españolas las muestras de clinker suministrada, asi como a los miembros del Comité Técnico "180-QUIC" (Qualitative Identification of Clínkers and Cements) de la RILEM (Réunion International des Laboratoires d¨Essais et de Recherches sur les Matériaux et les constructions). Janos Abonyi agradece la beca de investigación Janos Bolyai de la Academia de Ciencias de Hungría y la ayuda del Ministerio de Educación de Hungría, FKFP 0073/2001. F. Puertas agradece a B. Torroja y a R. Torroja su inestimable ayuda en la traducción y revisión técnica de la parte estadística.

The results show that the proposed method is useful to identify compact classifiers that are able to determine the origin of the clinker. A detailed description helps the implementation of the decision tree induction algorithm; still easier, a program has been constructed and can be downloaded (www.fmt.vein.hu/softcomp).

6. ACKNOWLEDGEMENTS

The financial support of OTKA (Hungarian National Research Foundation), No. T026307 is gratefully acknowledged. Thanks to spanish cement factories for collecting the clinker and also to 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). Janos Abonyi is grateful for the Janos Bolyai Research Fellowship of the Hungarian Academy of Science and for the support of Hungarian Ministry of Education, FKFP 0073/2001. F. Puertas thanks to B. Torroja and R. Torroja for their inestimable helpful for the translation and the technical revision of statistical part

BIBLIOGRAFÍA

(1) R.L. Goguel, D.A. St. John. Chemical identification of Portland cements in New Zealand concretes, Part I. Characteristic differences among New Zealand cements in minor and trace element chemistry. Cem Concr Res 23 (1) (1993) 59-68; Part II. The Ca-Sr-Mn plot in cement identification and the effect of aggregates. Cem Concr Res 23 (2) (1993) 283-293.
(2) J. C. Miller. J. N. Miller. Statistics for analytical chemistry; chapter 7.13: Pattern recognition, Ellis Horwood Ltd., New York, 1984.
(3) F.D. Tamas. Pattern recognition methods for the qualitative identification of Hungarian clinkers. World Cement / Res. & Development 27 (1996) 75-79.
(4) F. D. Tamás, É. Kristóf-Makó. Chemical "fingerprints" in Portland cement clinkers in: A. Gerdes (Ed.), Advances in Building Materials Science–Festschrift Wittmann, Aedificatio Publishers, Freiburg – Unterengstringen, 1996, pp. 217-228.
(5) F. D. Tamás, A. Tagnit-Hamou, J. Tritthart. Trace elements in clinker and their use as "fingerprints" to facilitate their qualitative identification. In: M. Cohen, S. Mindess, J. Skalny (Eds.), Materials Science of Concrete–The Sidney Diamond Symposium, Honolulu, HI, September 1998. American Ceramic Society, Westerville OH, pp. 57-69
(6) F. D. Tamás, J. Abonyi. Trace elements in clinkers – I. A graphical representation. Cem Concr Res, submitted for publication.
(7) F. D. Tamás, J. Abonyi. Trace elements in clinkers – II. Qualitative identification by fuzzy clustering. Cem Concr Res, submitted for publication.
(8) Dennis H. Rouvray (Editor). Fuzzy Logic in Chemistry, Academic Press, 1997.
(9) G. Barkó, J. Abonyi, J. Hlavay. Application of Fuzzy Clustering and Piezoelectric Chemical Sensor Array for Investigation on Organic Compounds, Analitica Chimica Acta, 398 (2-3), 219-22, 1999.
(10) J. R. Quinlan. C4.5: Programs for Machine Learning, Morgan Kaufman, San Mateo, 1993.
(11) J. R. Quinlan. Improved Use of Continuous Attributes in C4.5, Journal of Artificial Intelligence Research, 4, 77-90, 1996.

* * *