AN EXPERIENCE IN MANAGEMENT OF IMPRECISE SOIL DATABASES BY MEANS OF FUZZY ASSOCIATION RULES AND FUZZY APPROXIMATE DEPENDENCIES [1]

J. Calero, G. Delgado, M. Sánchez-Marañón
Department of Pedology and Agricultural Chemistry. University of Granada
{*varanda,gdelgado,masanche*}*@ugr.es*

D. Sánchez, M.A.Vila
Department of Computer Science and A.I. University of Granada
{*daniel,vila*}*@decsai.ugr.es*

J.M. Serrano
Department of Computer Science. University of Jaen
jschica@ujaen.es

Keywords: Expert soil knowledge, aggregated soil databases, imprecision factors in soil knowledge, fuzzy data mining.

Abstract: In this work, we start from a database built with soil information from heterogeneous scientific sources (Local Soil Databases, LSDB). We call this an Aggregated Soil Database (ASDB). We are interested in determining if knowledge obtained by means of fuzzy association rules or fuzzy approximate dependencies can represent adequately expert knowledge for a soil scientific, familiarized with the study zone. A master relation between two soil attributes was selected and studied by the expert, in both ASDB and LSDB. Obtained results reveal that knowledge extracted by means of fuzzy data mining tools is significatively better than crisp one. Moreover, it is highly satisfactory from the soil scientific expert's point of view, since it manages with more flexibility imprecision factors (IFASDB) commonly related to this type of information.

1 INTRODUCTION

Soil survey data is required for different kinds of environmental and agronomic studies, specially for estimation of soil quality indicators and other very important soil characteristics over large areas (Cazemier et al., 2001). Many of these parameters present a high degree of spatial variability, and they obstruct knowledge extraction when soil survey scale is small or very small (1:200000 or lower). In other order of things, obtaining a high precision map can be very expensive in time and resources, as a minimum number of measures would be desirable for resource optimization. Due to costs related to the schedule of a cartography or soil survey at a high scale in large geographic areas, researchers must recur in many occasions to knowledge fusion from different local soil databases for regional or national level studies (Bui and Moran, 2003).

Information sources in local soil databases present a very heterogeneous nature, combining not only soil cartographies but also Ph.D. thesis, monographes and other diverse works. This fact implies that resulting databases from local soil databases fusion (Aggregated soil databases, ASDB) present an additional imprecision or uncertainty degree related to local information aggregation processes.

Statistical analysis techniques are frequently applied in soil study: analysis of variance (Ulery and Graham, 1993), regression analysis (Qian et al., 1993), main components analysis (Sánchez-Marañón et al., 1996) and discriminant analysis (Scheinost and Schwertmann, 1999). These techniques, based on statistical probability theory, are adequate for dealing with uncertainty derived from randomness. Nevertheless, they are not suitable when managing imprecision or uncertainty related to qualitative character in many attributes (soil structure, consistency), of subjective nature and hard for mathematical treatment (Webster, 1977), as the ones in the ASDB.

Data mining techniques (such as association rules or approximate dependencies) have been proven as effective tools when looking for hidden or implicit relations between attributes in a large database (ASDB) and they do not have the limitations of statistical procedures commented above. In particular, fuzzy data mining tools can be specially suitable when we consider intrinsically fuzzy information, as soil data.

In this work, our objective is to extract knowledge from an ASDB obtained from local heterogeneous information sources. We want to test that fuzzy data

[1]This work is supported by the research project FuzzyKIM, CICYT TIC2002-04021-C02-02.

I. Seruca et al. (eds.), Enterprise Information Systems VI, 158–166.

mining tools can manage the increment in imprecision or uncertainty degree related to an aggregation process, better than crisp tools. In order to accomplish this, we introduce a methodology of knowledge extraction and interpretation on an ASDB real case, in a large area in Iberian Peninsula Southeast. From this, the domain expert will estimate the suitability of the proposed tools for this particularly difficult case of databases.

2 PROBLEM STATEMENT

We consider several soil databases with an analogous structure, but obtained from different sources, by means of different criteria. Modelling these sources and criteria as discriminant attributes, we can fuse all these databases into an ASDB. Also, in order to model soil data more accurately, we consider fuzzy similarity relations and sets of linguistic labels over some soil attributes.

Formally, let $RE = \{A_1, \ldots, A_m, B\}$ be a relational scheme, and r an instance of RE. Let $S_{RE} = \{S_{A_k}\}$ be a set of fuzzy similarity relations over attributes in RE. Also, given $t \in r$ a tuple, let $t[A_k]$ be the intersection of t and A_k, and $\mu_{t[A_k]}$ the membership degree of the value.

Finally, let B be a discriminant attribute, and $dom(B) = \{b_1, \ldots, b_l\}$ the set of possible values of B. Our idea when using discriminant attributes is to generate more homogenous sets of objects from a database, according to the attribute values. That is, we can perform a query as the following,

$$\texttt{select } A_1, \ldots, A_m \texttt{ from } r \texttt{ where } B = b_j;$$

obtaining r_j, a subset of r. Each subrelation r_j can be viewed as a LSDB, obtained according to a given criterium.

Moreover, if we apply some data mining techniques (i.e., association rules or approximate dependencies, see below), we can obtain the following,

- R_G, the set of all rules from r.
- R_j, the set of all rules from r_j.

We are interested in the study of possible existing correspondences between these sets of rules:

- Can we find rules in R_j that do not hold in R_G, and viceversa?
- Can imprecision or uncertainty management in data generate more accurate rules from domain experts' point of view, at both levels R_G and R_j?

Our proposed methodology is the following:

1. To define a set of criteria for decomposition of a given ASDB (r) into several LSDB (r_j), according to B, discriminant attribute, values.
2. To extract (fuzzy) approximate dependencies between attributes in r and in every subset of r.
3. To describe the obtained dependencies at a local level, by means of (fuzzy) association rules.
4. To compare the resulting sets of rules and dependencies in order to discover possible couplings at different levels.
5. To study in which real world problems imprecision and uncertainty management in data can generate better rules or dependencies, that is, when domain experts find more interesting and reasonable the obtained results.

3 DATA MINING TOOLS

In this section we summarize the techniques we have employed to analyze data corresponding to soil color and properties.

3.1 Association Rules

Given a set I ("set of items") and a set of transactions T (also called T-set), each transaction being a subset of I, association rules are "implications" of the form $A \Rightarrow C$ that relate the presence of itemsets (sets of items) A and C in transactions of T, assuming $A, C \subseteq I$, $A \cap C = \emptyset$ and $A, C \neq \emptyset$.

In the case of relational databases, it is usual to consider that items are pairs $\langle attribute, value \rangle$, and transactions are tuples in a table. For example, the item $\langle X, x_0 \rangle$ is in the transaction associated to a tuple t iff $t[X] = x_0$.

The ordinary measures proposed in (Agrawal et al., 1993) to assess association rules are *confidence* (the conditional probability $p(C|A)$) and *support* (the joint probability $p(A \cup C)$).

An alternative framework was proposed in (Berzal et al., 2001; Berzal et al., 2002). In this framework, accuracy is measured by means of Shortliffe and Buchanan's certainty factors (Shortliffe and Buchanan, 1975), in the following way: the certainty factor of the rule $A \Rightarrow C$ is

$$CF(A \Rightarrow C) = \frac{(Conf(A \Rightarrow C)) - S(C)}{1 - S(C)} \quad (1)$$

if $Conf(A \Rightarrow C) > S(C)$, and

$$CF(A \Rightarrow C) = \frac{(Conf(A \Rightarrow C)) - S(C)}{S(C)} \quad (2)$$

if $Conf(A \Rightarrow C) < S(C)$, and 0 otherwise.

Certainty factors take values in $[-1, 1]$, indicating the extent to which our belief that the consequent is true varies when the antecedent is also true. It ranges

from 1, meaning maximum increment (i.e., when A is true then C is true) to -1, meaning maximum decrement.

3.2 Approximate Dependencies

A functional dependence $V \rightarrow W$ holds in a relational scheme RE if and only if $V, W \subseteq RE$ and for every instance r of RE

$$\forall t, s \in r \;\; if\; t[V] = s[V]\; then\; t[W] = s[W] \quad (3)$$

Approximate dependencies can be roughly defined as functional dependencies with exceptions. The definition of approximate dependence is then a matter of how to define exceptions, and how to measure the accuracy of the dependence (Bosc and Lietard, 1997). We shall follow the approach introduced in (Delgado et al., 2000; Blanco et al., 2000), where the same methodology employed in mining for AR's is applied to the discovery of AD's.

The idea is that, since a functional dependency "$V \rightarrow W$" can be seen as a rule that relates the equality of attribute values in pairs of tuples (see equation (3)), and association rules relate the presence of items in transactions, we can represent approximate dependencies as association rules by using the following interpretations of the concepts of item and transaction:

- An item is an object associated to an attribute of RE. For every attribute $At_k \in RE$ we note it_{At_k} the associated item.
- We introduce the itemset I_V to be
$$I_V = \{it_{At_k} \mid At_k \in V\}$$
- T_r is a T-set that, for each pair of tuples $< t, s > \in r \times r$ contains a transaction $ts \in T_r$ verifying
$$it_{At_k} \in ts \;\Leftrightarrow\; t[At_k] = s[At_k]$$
It is obvious that $|T_r| = |r \times r| = n^2$.

Then, an approximate dependence $V \rightarrow W$ in the relation r is an association rule $I_V \Rightarrow I_W$ in T_r (Delgado et al., 2000; Blanco et al., 2000). The support and certainty factor of $I_V \Rightarrow I_W$ measure the interest and accuracy of the dependence $V \rightarrow W$.

3.3 Fuzzy Association Rules

In (Delgado et al., 2003), the model for association rules is extended in order to manage fuzzy values in databases. The approach is based on the definition of fuzzy transactions as fuzzy subsets of items. Let $I = \{i_1, \ldots, i_m\}$ be a set of items and T' be a set of fuzzy transactions, where each fuzzy transaction is a fuzzy subset of I. Let $\tilde{\tau} \in T'$ be a fuzzy transaction, we note $\tilde{\tau}(i_k)$ the membership degree of i_k in $\tilde{\tau}$. A fuzzy association rule is an implication of the form $A \Rightarrow C$ such that $A, C \subset RE$ and $A \cap C = \emptyset$.

It is immediate that the set of transactions where a given item appears is a fuzzy set. We call it *representation* of the item. For item i_k in T' we have the following fuzzy subset of T':

$$\tilde{\Gamma}_{i_k} = \sum_{\tilde{\tau} \in T'} \tilde{\tau}(i_k)/\tilde{\tau} \quad (4)$$

This representation can be extended to itemsets as follows: let $I_0 \subset I$ be an itemset, its representation is the following subset of T':

$$\tilde{\Gamma}_{I_0} = \bigcap_{i \in I_0} \tilde{\Gamma}_i = min_{i \in I_0} \tilde{\Gamma}_i \quad (5)$$

In order to measure the interest and accuracy of a fuzzy association rule, we must use approximate reasoning tools, because of the imprecision that affects fuzzy transactions and, consequently, the representation of itemsets. In (Delgado et al., 2003), a semantic approach is proposed based on the evaluation of quantified sentences (see (Zadeh, 1983)). Let Q be a fuzzy coherent quantifier:

- The support of an itemset $\tilde{\Gamma}_{I_0}$ is equal to the result of evaluating the quantified sentence Q of T' are $\tilde{\Gamma}_{I_0}$.
- The support of the fuzzy association rule $A \Rightarrow C$ in the FT-set T', $Supp(A \Rightarrow C)$, is the evaluation of the quantified sentence Q of T are $\tilde{\Gamma}_{A \cup C}$ = Q of T are $(\tilde{\Gamma}_A \cap \tilde{\Gamma}_C)$.
- The confidence of the fuzzy association rule $A \Rightarrow C$ in the FT-set T', $Conf(A \Rightarrow C)$, is the evaluation of the quantified sentence Q of $\tilde{\Gamma}_A$ are $\tilde{\Gamma}_C$.

As seen in (Delgado et al., 2003), the proposed method is a generalization of the ordinary association rule assessment framework in the crisp case.

3.4 Fuzzy Approximate Dependencies

As seen in (Bosc and Lietard, 1997), it is possible to extend the concept of functional dependence in several ways by smoothing some of the elements of the rule in equation 3. We want to consider as much cases as we can, integrating both approximate dependencies (exceptions) and fuzzy dependencies. For that purpose, in addition to allowing exceptions, we have considered the relaxation of several elements of the definition of functional dependencies. In particular we consider membership degrees associated to pairs (attribute, value) as in the case of fuzzy association rules, and also fuzzy similarity relations to smooth the equality of the rule in equation 3.

We shall define fuzzy approximate dependencies in a relation as fuzzy association rules on a special FT-set obtained from that relation, in the same way that approximate dependencies are defined as association rules on a special T-set.

Let $I_{RE} = \{it_{At_k} | At_k \in RE\}$ be the set of items associated to the set of attributes RE. We define a FT-set T'_r associated to table r with attributes in RE as follows: for each pair of rows $< t, s >$ in $r \times r$ we have a fuzzy transaction ts in T'_r defined as

$$\forall it_{At_k} \in T'_r, ts(it_{At_k}) =$$

$$min(\mu_t(At_k), \mu_s(At_k), S_{At_k}(t(At_k), s(At_k))) \quad (6)$$

This way, the membership degree of a certain item it_{At_k} in the transaction associated to tuples t and s takes into account the membership degree of the value of At_k in each tuple and the similarity between them. This value represents the degree to which tuples t and s agree in At_k, i.e., the kind of items that are related by the rule in equation 3. On this basis, we define fuzzy approximate dependencies as follows (Berzal et al., 2003; Serrano, 2003):

Let $X, Y \subseteq RE$ with $X \cap Y = \emptyset$ and $X, Y \neq \emptyset$. The fuzzy approximate dependence $X \rightarrow Y$ in r is defined as the fuzzy association rule $I_X \Rightarrow I_Y$ in T'_r.

The support and certainty factor of $I_X \Rightarrow I_Y$ are calculated from T'_r as explained in sections 3.3 and 3.1, and they are employed to measure the importance and accuracy of $X \rightarrow Y$.

A FAD $X \rightarrow Y$ holds with total accuracy (certainty factor $CF(X \rightarrow Y) = 1$) in a relation r iff $ts(I_X) \leq ts(I_Y)\ \forall ts \in T'_r$ (let us remember that $ts(I_X) = \min_{At_k \in X} ts(it_{At_k})\ \forall X \subseteq RE$). Moreover, since fuzzy association rules generalize crisp association rules, FAD's generalize AD's.

Additional properties and an efficient algorithm for computing FAD's can be found in (Berzal et al., 2003; Serrano, 2003).

3.5 Fuzzy Association Rules with Fuzzy Similarity Relations

Fuzzy logic can be an effective tool for representation of heterogeneous data. In fact, fuzzy similarity relations allow us to establish semantic links between values.

Several fuzzy association rules definitions can be found in the literature but, to our knowledge, none of them contemplates fuzzy similarity relations between values. Given two items $i_0 =< A, a_0 >$ and $i_1 =< A, a_1 >$, and a similarity degree $S_A(a_0, a_1) = \alpha$, it would be desirable to have into account how the support of an item is affected by appearances of similar items.

In (Sánchez et al., 2004), we extend the definition of fuzzy association rule (section 3.3)in the following way. Let $A \in RE$ be an attribute, and $dom(A) = \{a_1, \ldots, a_p\}$ the set of possible values of A. For each $a_i \in A$, we define a linguistic label E_{a_i} as the function

$$E_{a_i} : A \rightarrow [0, 1]; E_{a_i}(a) = S_A(a_i, a) \quad (7)$$

where $S_A(a_i, a)$ is the similarity degree between a_i and a. Let I_A be the set of items where each item is associated to a pair $< A, E_{a_i} >$, $|I_A| = |dom(A)|$. This way, each time an item appears, we reflect its similarity with other items as the compatibility degree returned by its linguistic label. Moreover, according to this representation, we can apply the same methodology proposed in (Delgado et al., 2003) in order to obtain fuzzy association rules.

4 EXPERIMENTS

To carry out the aggregation process, we started from 14 databases, created from local information sources, that constitute the so called Local Soil Databases (LSDB). In this context, we denominated *"local"* information source each one of the categories for Discriminant Attributes in Table 1. Likewise, the Aggregated Soil Database (ASDB) results from the *"aggregation"* or inclusion in one large database of every local information source. During this process, a number of factors, that we called imprecision factors in Aggregated Soil Databases (IFASDB), appeared, causing a loss of accuracy and effectiveness in representation, extraction and management of knowledge allusive to the problem in the real world at ASDB level. We could describe several IFASDB, but in this work we considered only three that resume, in great part, all the others. This factors are: the ecogeographical variability, the bibliography from we extracted data and the set of protocols and standard techniques used by authors to describe and analyze soils (discriminant attributes *Mesoenvironment*, *Bibliographic Source* and *Protocol*, respectively, in Table 1). At this point, we must also describe the mesoenvironments (Sierra Nevada, Sierra of Gádor and Southeast). Relations between soil attributes and values that can be studied by means of our data mining techniques are very numerous. The expert can enumerate a huge amount of basic well-known relations in Soil Science, i.e: mean annual rainfall and altitude, % of slope and % of clay, % of $CaCO_2$ and pH, original material and effective soil thickness, structure type and Horizon type, etc. We called all this rules *A Priori Expert Rules* (PER). From the set of PERs, we selected the rules derived from the dependence

$$HorizonType \rightarrow \%OrganicCarbon$$

This relates two very meaningful attributes in Soil Science:

- The horizon type definition and classification are conditioned for OC (Organic Carbon) content, a diagnostic feature in most employed systems of soil classification (Soil-Survey-Staff, 1975; FAO, 1998).
- OC content is highly sensitive to ecological and geographical variability in Mediterranean climate type.
- Both attributes are good indicators for several soil forming processes as melanization, accumulation, vertisolation, isohumification, horizonation, mineralization. . .
- OC content is an useful index for physical and biochemical degradation of soils, and it is in strict dependence with management.
- Analytical methods for OC content determination are not very sensitive to uncertainty, as opposed to the type of horizon. The latter is highly imprecise and is closely related with the analyst's competence and finality.

Once PERs are selected, we study the obtained ARs, ADs, FARs and FADs at both local and aggregated levels (LSDB and ASDB, respectively). By means of CF, we assess the extracted knowledge and suggest appropriate considerations for use of this data mining techniques, from an expert's point of view.

4.1 Knowledge Sources. Pretreatment of Soil Information

The ASDB included soil information about three mesoenvironments from the South and Southeast of the Iberian Peninsula under Mediterranean climate: Sierra Nevada, Sierra of Gádor and Southeast (involving part of the provinces of Murcia and Almería). Table 1 shows the main characteristics of local information sources. We used two Ph.D. Thesis and five cartographic sheets from LUCDEME, scale 1:100000.

Data from Sierra of Gádor was extracted from (Oyonarte, 1990) and consists of 70 soil profiles and 176 horizons. Altitude fluctuates from 100 to 2200 m, and rainfall from 213 mm/year (semiarid climate) to 813 mm/year (wet climate), with a mean annual rainfall of 562 mm/year. Lowest annual mean temperature is 6.4 C and the highest is 21.0 C, with a mean of 12.7 C. Original soil materials are of carbonated type, mainly limestones and dolomites.

Data from Southeast was extracted from LUCDEME soil maps, specifically from sheets 1041 from Vera, Almería (Delgado et al., 1991), 911 from Cehegin, Murcia (Alias, 1987), 1030 from Tabernas, Almería (Pérez Pujalte, 1987), 912 from Mula, Murcia (Alias, 1986) and 1031 from Sorbas, Almería (Pérez Pujalte, 1989). There is a total of 89 soil profiles and 262 horizons. Altitude fluctuates from 65 to 1120 m, and rainfall from 183 mm/year (arid climate) to 359 mm/year (semiarid climate), with a mean annual rainfall of 300 mm/year. Lowest annual mean temperature is 13.2 C and the highest is 19.0 C, with a mean of 17.0 C. Geological environment and Original soil materials are extremely different, we can find carbonated, acids and volcanic rocks.

Data from Sierra Nevada was extracted from (Sánchez-Marañón, 1992). There is a total of 35 soil profiles and 103 horizons. Altitude fluctuates from 1320 to 3020 m, and rainfall from 748 mm/year (semihumid climate) to 1287 mm/year (hiperhumid climate), with a mean annual rainfall of 953 mm/year. Lowest annual mean temperature is 0.0 C and the highest is 12.1 C. Geological environment and Original soil materials are mainly acids, but it is not strange to find basic rocks.

Attributes with numeric domains were discretized, following some of the methodologies discussed in (Hussain et al., 1999), under supervision of domain experts. A set of linguistic labels $\{Low, Medium, High\}$ was defined for every numeric attribute. Attributes with categorical domains were fuzzified considering fuzzy similarity relations.

4.2 Analyzing Discovered Knowledge

4.2.1 Crisp Case

When we considered crisp relations from ASDB (Table 2), we found only one AD, $HorizonType \rightarrow \%OrganicCarbon$ with CF 0.089, that reveal a strong grade of independence between these attributes. Provisionally, this conclusion contradicts the expert experience, confirmed in the bibliography. As we could expect, we obtained only four ARs, mainly with consequent $[\%OrganicCarbon = Low]$. This fact was not surprising to us, because the *"Low"* category had a high support (70%) ASDB. As the support threshold for rules was 10%, rules having *"Medium"* and *"High"* categories, were not found. In both cases, crisp data mining was not satisfactory enough for Soil Scientists, and we could not *"fuse"* ASDB and expert knowledge. Otherwise, when we considered separately the information stored in LSDBs (Table 3), we obtained approximate dependencies with higher CF than in ASDB. This phenomenon could reflect the action of IFASDB. Despite of this, some local dependencies showed smaller CF values than in the aggregated case, and express a total independence.

4.2.2 Fuzzy Case

Observing FADs from ASDB, a CF of 0.31 is found (Table 2). Despite of this low CF, the dependence degree shown between $HorizonType$ and OC content was more informative than in the crisp case. It reflected better the expert knowledge. Even though, initially, the soil scientist expected a higher degree, it can be explained due to the influence of soils placed at Southeast Mesoenvironment in ASDB. Indeed, due to the arid nature of this climate, it could be expected that one of the main soil forming factors, OC content incorporated to soil from vegetation, were low and homogenous. The latter conclusion can be checked regarding Table 4. Moreover, fuzzy data mining let us obtain a higher number of rules than in crisp case. This supposes, quantitatively, a higher volume of discovered knowledge.

A good example of correspondence or *"fusion"* between databases and expert knowledge could be obtained comparing ARs from Sierra of Gádor with Southeast ones. The former had rules with *"moderate"* and *"high"* OC content in consequent, whereas the latter had a *"low"* value in consequent. Sierra of Gádor has a higher mean altitude and mean annual rainfall, and, consequently, more vegetation in soil and horizons (especially in *Ah* type). Looking at this, the fuzzy model reflects more accurately soil forming processes as melanization and accumulation. We can also examine others IFSDB in addition to *Mesoenvironment*. I.e., *Protocol* constitute an important source of variability in ASDB. Comparing *"Perez"* and *"Alias"* categories, the former has more ARs (Table 6) and relates more categories, reflecting a more detailed and precise knowledge than *"Alias"*. *"Perez"* protocols (including field description, analysis and other techniques) seem to be more reliable than *"Alias"* ones.

5 CONCLUSIONS

We have seen how large databases can be divided into homogeneous subsets defining one or more discriminant attributes. This division, followed by a knowledge discovery process, can allow us to discover previously unnoticed relations in data.

We conclude that, for this particular case, knowledge extracted by means of fuzzy data mining was more suitable to *"fusion"* or comparison with expert knowledge that crisp. Moreover, fuzzy data mining was sensitive to low support categories as $[\%OrganicCarbon = Low]$ or $[HorizonType = Bk\ or\ Btk]$, discarded in crisp data mining.

We could confirm that fuzzy data mining is highly sensitive to latent knowledge in ASDBs. That fact is very important for a soil scientist, since lets us apply it with the assurance that imprecision and uncertainty factors (IFASDB) will not distort or alter the knowledge discovery process.

As a future task, we propose to solve this same problem in a general case. With a domain expert aid, we must define the set of criteria for database decomposition but also discern when fuzzy techniques get better results than crisp ones.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proc. Of the 1993 ACM SIGMOD Conference*, pages 207–216.

Alias, J. (1986). *Mapa de suelos de Mula. Mapa 1:100000 y memoria*. LUCDEME; MAPA-ICONA-University of Murcia.

Alias, J. (1987). *Mapa de suelos de Cehegin. Mapa 1:100000 y memoria*. LUCDEME; MAPA-ICONA-University of Murcia.

Berzal, F., Blanco, I., Sánchez, D., Serrano, J., , and Vila, M. (2003). A definition for fuzzy approximate dependencies. *Fuzzy Sets and Systems*. Submitted.

Berzal, F., Blanco, I., Sánchez, D., and Vila, M. (2001). A new framework to assess association rules. In Hoffmann, F., editor, Advances in Intelligent Data Analysis. Fourth International Symposium, IDA'01. Lecture Notes in Computer Science 2189, pages 95–104. Springer-Verlag.

Berzal, F., Blanco, I., Sánchez, D., and Vila, M. (2002). Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*. An extension of (Berzal et al., 2001), submitted.

Blanco, I., Martín-Bautista, M., Sánchez, D., and Vila, M. (2000). On the support of dependencies in relational databases: strong approximate dependencies. *Data Mining and Knowledge Discovery*. Submitted.

Bosc, P. and Lietard, L. (1997). Functional dependencies revisited under graduality and imprecision. In Annual Meeting of NAFIPS, pages 57–62.

Bui, E. and Moran, C. (2003). A strategy to fill gaps in soil over large spatial extents: An example from the murray-darlin basin of australia. *Geoderma, 111*, pages 21–44.

Cazemier, D., Lagacherie, P., and R., M.-C. (2001). A possibility theory approach from estimating available water capacity from imprecise information contained in soil databases. *Geoderma, 103*, pages 113–132.

Delgado, G., Delgado, R., Gamiz, E., Párraga, J., Sánchez Marañon, M., Medina, J., and Martín-García, J. (1991). *Mapa de Suelos de Vera*. LUCDEME, ICONA-Universidad de Granada.

Delgado, M., Marín, N., Sánchez, D., and Vila, M. (2003). Fuzzy association rules: General model and applications. *IEEE Transactions on Fuzzy Systems 11(2)*, pages 214–225.

Delgado, M., Martín-Bautista, M., Sánchez, D., and Vila, M. (2000). Mining strong approximate dependencies from relational databases. In Proceedings of IPMU'2000.

FAO (1998). The world reference base for soil resources. world soil resources. Technical Report 84, ISSS/AISS/IBG/ISRIC/FAO, Rome.

Hussain, F., Liu, H., Tan, C., and Dash, M. (1999). Discretization: An enabling technique. Technical report, The National University of Singapore.

Oyonarte, C. (1990). *Estudio Edáfico de la Sierra de Gádor (Almera). Evaluación para usos forestales*. PhD thesis, University of Granada.

Pérez Pujalte, A. (1987). *Mapa de suelos de Tabernas. Mapa 1:100000 y memoria*. LUCDEME; MAPA-ICONA-CSIC.

Pérez Pujalte, A. (1989). *Mapa de suelos de Sorbas. Mapa 1:100000 y memoria*. LUCDEME; MAPA-ICONA-CSIC.

Qian, H., Klinka, K., and Lavkulich, L. (1993). Relationships between color value and nitrogen in forest mineral soils. *Can. J. Soil Sci., 73*, pages 61–72.

Sánchez, D., Sánchez, J. R., Serrano, J. M., and Vila, M. A. (2004). Association rules over imprecise domains involving fuzzy similarity relations. To be submitted to Estylf 2004.

Sánchez-Marañón, M. (1992). *Los suelos del Macizo de Sierra Nevada. Evaluación y capacidad de uso (in Spanish)*. PhD thesis, University of Granada.

Sánchez-Marañón, M., Delgado, R., Párraga, J., and Delgado, G. (1996). Multivariate analysis in the quantitative evaluation of soils for reforestation in the sierra nevada (southern spain). *Geoderma, 69*, pages 233–248.

Scheinost, A. and Schwertmann, U. (1999). Color identification of iron oxides and hydroxisulfates: Uses and limitations. *Soil Sci. Soc. Am. J., 65*, pages 1463–1461.

Serrano, J. (2003). *Fusin de Conocimiento en Bases de Datos Relacionales: Medidas de Agregacin y Resumen (in Spanish)*. PhD thesis, University of Granada.

Shortliffe, E. and Buchanan, B. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, *23*, pages 351–379.

Soil-Survey-Staff (1975). *Soil Taxonomy*. U.S. Dept. Agri. Handbook No. 436.

Ulery, A. and Graham, R. (1993). Forest-fire effects on soil color and texture. *Soil Sci. Soc. Am. J., 57*, pages 135–140.

Webster, R. (1977). *Quantitative and Numerical Methods in Soil Classification and Survey*. Claredon Press, Oxford.

Zadeh, L. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computing and Mathematics with Applications*, 9(1):149–184.

Table 1: Discriminant attributes for the soil database

Mesoenvironment	**Soil profile**	**Horizon**
Sierra Nevada [SN]	35	103
Sierra Gádor[SB]	70	176
Southeast [SE]	89	262
Bibliographic source	**Soil profile**	**Horizon**
Ph.D. Thesis [MARAÑON]	35	103
Ph.D. Thesis [OYONARTE]	70	176
LUCDEME sheet 1014, Vera.[VERA]	29	76
LUCDEME sheet 1030, Tabernas. [TABERNA]	14	37
LUCDEME sheet 1031, Sorbas. [SORBAS]	24	72
LUCDEME sheet 912, Cehegn. [CEHEGIN]	10	32
LUCDEME sheet 911, Mula. [MULA]	12	45
Acting protocol	**Soil profile**	**Horizon**
Sánchez-Marañon, M. [SANCHEZ]	35	103
Oyonarte, C. [CECILIO]	70	176
Pérez-Pujalte, A. [PEREZ]	67	185
Alías, L. [ALIAS]	22	77

Table 2: Obtained CF in ASDB ($HorizonType \rightarrow \%OrganicCarbon$)

Approx. Dep.	CF 0.09
Assoc. Rules	$[C] \Rightarrow [L]$, CF 0.8 $[Bw] \Rightarrow [L]$, CF 0.7 $[Ah] \Rightarrow [L]$, CF -0.39 $[Ah] \Rightarrow [M]$, CF 0.41
F. Approx. Dep.	CF 0.31
F. Assoc. Rules	$[Ck] \Rightarrow [L]$, CF 0.53 $[C] \Rightarrow [L]$, CF 0.69 $[Bwk] \Rightarrow [L]$, CF 0.23 $[Bw] \Rightarrow [L]$, CF 0.41 $[Bw] \Rightarrow [M]$, CF 0.25 $[Btk] \Rightarrow [L]$, CF 0.81 $[Bt] \Rightarrow [L]$, CF 1 $[Bk] \Rightarrow [L]$, CF 0.49 $[Ap] \Rightarrow [L]$, CF 0.50 $[Ah] \Rightarrow [L]$, CF -0.01 $[Ah] \Rightarrow [M]$, CF 0.13 $[Ah] \Rightarrow [H]$, CF 0.28

Table 3: Obtained CF in crisp LSDB ($HorizonType \rightarrow \%OrganicCarbon$)

Mesoenv.	**AD**	**AR**
SE	CF 0.01	$[Ap] \Rightarrow [L]$, CF 0.56 $[Bw] \Rightarrow [L]$, CF 0.81 $[Ah] \Rightarrow [L]$, CF -0.19 $[Ah] \Rightarrow [M]$, CF 0.23
SB	CF 0.37	$[Bw] \Rightarrow [L]$, CF 0.65 $[Ah] \Rightarrow [L]$, CF -0.58 $[Ah] \Rightarrow [M]$, CF 0.57
SN	CF 0.11	$[C] \Rightarrow [L]$, CF 1 $[Bw] \Rightarrow [L]$, CF 1 $[Ah] \Rightarrow [L]$, CF -0.35 $[Ah] \Rightarrow [M]$, CF 0.35
Bib. source	**AD**	**AR**
CEHEGIN	CF 0.12	$[Ck] \Rightarrow [L]$, CF -0.15 $[Ap] \Rightarrow [L]$, CF 1 $[Ah] \Rightarrow [L]$, CF -0.32
MARAÑON	CF 0.11	$[C] \Rightarrow [L]$, CF 1 $[Bw] \Rightarrow [L]$, CF 1 $[Ah] \Rightarrow [L]$, CF -0.35 $[Ah] \Rightarrow [M]$, CF 0.35
MULA	CF 0.73	$[C] \Rightarrow [L]$, CF 1 $[Ap] \Rightarrow [L]$, CF 1
OYONARTE	CF 0.37	$[Bw] \Rightarrow [L]$, CF 0.65 $[Ah] \Rightarrow [L]$, CF -0.58 $[Ah] \Rightarrow [M]$, CF 0.57
SORBAS	CF -0.01	$[C] \Rightarrow [L]$, CF 1 $[Bw] \Rightarrow [L]$, CF 1 $[Ap] \Rightarrow [L]$, CF -0.07 $[Ah] \Rightarrow [M]$, CF -0.02
TABERNAS	CF -0.02	$[C] \Rightarrow [L]$, CF -0.03 $[Bw] \Rightarrow [L]$, CF 1 $[Ap] \Rightarrow [L]$, CF -0.13 $[Ah] \Rightarrow [L]$, CF -0.01
VERA	CF 0.07	$[C] \Rightarrow [L]$, CF 0.5 $[Ap] \Rightarrow [L]$, CF -0.23
Acting prot.	**AD**	**AR**
ALIAS	CF 0.53	$[Ck] \Rightarrow [L]$, CF -0,23 $[Ck] \Rightarrow [L]$, CF 1 $[Ck] \Rightarrow [L]$, CF 1
CECILIO	CF 0.37	$[Bw] \Rightarrow [L]$, CF 0.65 $[Ah] \Rightarrow [L]$, CF -0.58 $[Ah] \Rightarrow [M]$, CF 0.57
SANCHEZ	CF 0.11	$[C] \Rightarrow [L]$, CF 1 $[Bw] \Rightarrow [L]$, CF 1 $[Ah] \Rightarrow [L]$, CF -0.35 $[Ah] \Rightarrow [M]$, CF 0.35
PEREZ	CF -0.04	$[C] \Rightarrow [L]$, CF 0.41 $[Bw] \Rightarrow [L]$, CF 0.81 $[Ah] \Rightarrow [L]$, CF -0.16 $[Ah] \Rightarrow [M]$, CF 0.19

Table 4: Obtained CF in fuzzy LSDB ($HorizonType \rightarrow \%OrganicCarbon$) (i)

Mesoenv.	**FAD**	**FAR**
SE	CF 0.38	$[Ck] \Rightarrow [L]$, CF 0.54 $[C] \Rightarrow [L]$, CF 0.73 $[Bwk] \Rightarrow [L]$, CF 0.43 $[Bw] \Rightarrow [L]$, CF 0.80 $[Btk] \Rightarrow [L]$, CF 0.88 $[Bt] \Rightarrow [L]$, CF 0.61 $[Bk] \Rightarrow [L]$, CF -0.03 $[Ap] \Rightarrow [L]$, CF 0.66 $[Ah] \Rightarrow [L]$, CF 0.23 $[Ah] \Rightarrow [M]$, CF 0.18
SB	CF 0.35	$[C] \Rightarrow [M]$, CF -0.05 $[C] \Rightarrow [H]$, CF -0.05 $[Bwk] \Rightarrow [M]$, CF 0.57 $[Bwk] \Rightarrow [H]$, CF 0.06 $[Btk] \Rightarrow [M]$, CF -0.05 $[Btk] \Rightarrow [H]$, CF -0.05 $[Bw] \Rightarrow [M]$, CF 0.48 $[Bw] \Rightarrow [H]$, CF 0.11 $[Bt] \Rightarrow [M]$, CF 0.31 $[Bt] \Rightarrow [H]$, CF -0.05 $[Ap] \Rightarrow [M]$, CF 0.30 $[Ap] \Rightarrow [H]$, CF -0.05 $[Ah] \Rightarrow [M]$, CF 0.07 $[Ah] \Rightarrow [H]$, CF 0.37
SN	CF 0.34	$[Ck] \Rightarrow [L]$, CF 0.80 $[C] \Rightarrow [L]$, CF 0.72 $[Bwk] \Rightarrow [L]$, CF 0.81 $[Bw] \Rightarrow [L]$, CF 0.41 $[Bw] \Rightarrow [M]$, CF 0.30 $[Bt] \Rightarrow [L]$, CF -0.01 $[Ap] \Rightarrow [L]$, CF -0.01 $[Ah] \Rightarrow [L]$, CF 0.26 $[Ah] \Rightarrow [M]$, CF 0.15 $[Ah] \Rightarrow [H]$, CF 0.05

Table 5: Obtained CF in fuzzy LSDB ($HorizonType \rightarrow \%OrganicCarbon$) (ii)

Bib. source	FAD	FAR
CEHEGIN	CF 0.52	$[Ck] \Rightarrow [L]$, CF 0.55 $[C] \Rightarrow [L]$, CF 0.80 $[Bwk] \Rightarrow [L]$, CF 0.85 $[Bw] \Rightarrow [L]$, CF 0.85 $[Ap] \Rightarrow [L]$, CF 0.34 $[Ap] \Rightarrow [M]$, CF 0.26 $[Ah] \Rightarrow [L]$, CF 0.09 $[Ah] \Rightarrow [M]$, CF 0.48
MULA	CF 0.72	$[Ck] \Rightarrow [L]$, CF 0.77 $[C] \Rightarrow [L]$, CF 0.81 $[Bwk] \Rightarrow [L]$, CF -0.03 $[Bw] \Rightarrow [L]$, CF -0.03 $[Ap] \Rightarrow [L]$, CF 0.80 $[Ah] \Rightarrow [L]$, CF 0.33
SORBAS	CF 0.65	$[Ck] \Rightarrow [L]$, CF 0.96 $[C] \Rightarrow [L]$, CF 0.96 $[Bw] \Rightarrow [L]$, CF 0.95 $[Bt] \Rightarrow [L]$, CF 0.96 $[Ap] \Rightarrow [L]$, CF 0.85 $[Ah] \Rightarrow [L]$, CF 0.66
TABERNAS	CF 0.52	$[Ck] \Rightarrow [L]$, CF 0.93 $[C] \Rightarrow [L]$, CF 0.78 $[Bwk] \Rightarrow [L]$, CF 0.93 $[Bw] \Rightarrow [L]$, CF 0.91 $[Bt] \Rightarrow [L]$, CF 0.93 $[Ap] \Rightarrow [L]$, CF 0.72 $[Ah] \Rightarrow [M]$, CF 0.52
VERA	CF 0.25	$[Ck] \Rightarrow [M]$, **CF 0.79** $[C] \Rightarrow [L]$, CF 0.44 $[Bwk] \Rightarrow [L]$, **CF 0.77** $[Bw] \Rightarrow [L]$, CF 0.60 $[Btk] \Rightarrow [L]$, CF 0.78 $[Bt] \Rightarrow [L]$, CF 0.24 $[Bt] \Rightarrow [M]$, CF 0.37 $[Ap] \Rightarrow [L]$, CF 0.45 $[Ap] \Rightarrow [M]$, CF 0.14 $[Ah] \Rightarrow [L]$, CF 0.05 $[Ah] \Rightarrow [M]$, CF 0.16 $[Ah] \Rightarrow [H]$, CF 0.23

Table 6: Obtained CF in fuzzy LSDB ($HorizonType \rightarrow \%OrganicCarbon$) (iii)

Acting prot.	FAD	FAR
ALIAS	CF 0.61	$[Ck] \Rightarrow [L]$, CF 0.50 $[C] \Rightarrow [L]$, CF 0.85 $[Bwk] \Rightarrow [L]$, CF 0.44 $[Bw] \Rightarrow [L]$, CF 0.44 $[Ap] \Rightarrow [L]$, CF 0.62 $[Ah] \Rightarrow [L]$, CF 0.19 $[Ah] \Rightarrow [M]$, CF 0.33
PEREZ	CF 0.33	$[Ck] \Rightarrow [L]$, CF 0.57 $[C] \Rightarrow [L]$, CF 0.66 $[Bwk] \Rightarrow [L]$, CF 0.43 $[Bw] \Rightarrow [L]$, CF 0.82 $[Btk] \Rightarrow [L]$, CF 0.87 $[Bt] \Rightarrow [L]$, CF 0.60 $[Bk] \Rightarrow [L]$, CF -0.01 $[Ap] \Rightarrow [L]$, CF 0.67 $[Ah] \Rightarrow [L]$, CF 0.25 $[Ah] \Rightarrow [M]$, CF 0.15 $[Ah] \Rightarrow [H]$, CF 0.24