USING FUZZY DATACUBES IN THE STUDY OF TRADING STRATEGIES

M. Delgado Calvo-Flores, J. F. Nuñez Negrillo
Department of Computer Science and Artificial Intelligent, University of Granada

E. Gibaja Galindo
Department of Informatics and Numeric Analysis, University of Cordoba

C. Molina Fernández[1]
Department of Computer Science, University of Jaen

Keywords: Fuzzy OLAP, imprecision, exploratory analysis.

Abstract: A fuzzy multidimensional model can be used for exploratory analysis, modelling complex concepts that are very difficult to use in crisp ones. Some problems, as the edge problem, can be reduced using this approach. To hide the complexity of the fuzzy logic in this situation is important. In this paper we present an application of a fuzzy multidimensional model, that uses two layer representation to hide the complexity to the user, in the study of trading strategies.

1 INTRODUCTION

OLAP systems are exploratory analysis tools that are designed to work with a high amont of data in an efficiently way. Some statistical software include this kind of tools (e.g. SPSS has an analysis functionality based on a simple model of DataCubes).

Some concepts are not well modelled using crisp models (e.g. near the average value, a young company, etc.). Fuzzy logic has been widely used to model concepts in a more natural way to user. Using a fuzzy multidimensional model allows to apply OLAP using fuzzy concepts and to get more intuitive results. Nowadays, the use of data from different sources and the use of semi-structured (e.g. XML) and non-structured (e.g. plain text) sources is normal. Now the systems need to manage imprecision in the data and more flexible structures to represent the analysis domains. The fuzzy logic can help us to model this kind of imprecision. Some times we need to categorize continues values to reduce the complexity of the analysis. If we do it by dividing the range using crisp intervals the result can present the edge problem: two values very near belong to different intervals. If we relax the edge of the intervals we can reduce this problem. These relaxed intervals can be modelled if we use a fuzzy multidimensional model.

In this paper we propose to use a fuzzy multidimensional model to analyze data from the behavior of trading strategies and avoid the problems mentioned. The paper is organized as follow: next section is dedicated to present the main concepts of the fuzzy multidimensional model used and the OLAP system that implements it; section 3 presents the DataCube built and (Section 4) some examples of analysis over it. The main conclusions are collected in last section.

2 FUZZY MULTIDIMENSIONAL MODEL

In this section we briefly introduce the fuzzy multidimensional model. A more detailed description can be found in (Molina et al., 2006; Delgado et al., 2004). Here we only present the main concepts needed to understand the model implemented.

2.1 Fuzzy Multidimensional Structure

Definition 1 *A dimension is a tuple $d = (l, \leq_d, l_\perp, l_\top)$ where $l = l_i, i = 1, ..., n$ so that each l_i is a set of values*

[1] Corresponding author. Phone: +34 953 212 883.

Delgado Calvo-Flores M., F. Nuñez Negrillo J., Gibaja Galindo E. and Molina Fernández1 C. (2007).
USING FUZZY DATACUBES IN THE STUDY OF TRADING STRATEGIES.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - DISI*, pages 164-169
DOI: 10.5220/0002386001640169

$l_i = \{c_{i1},...,c_{in}\}$ and $l_i \cap l_j = \emptyset$ if $i \neq j$, and $\leq_d$ is a partial order relation between the elements of l so that $l_i \leq_d l_k$ if $\forall c_{ij} \in l_i \Rightarrow \exists c_{kp} \in l_k / c_{ij} \subseteq c_{kp}$. $l_\perp$ and $l_\top$ are two elements of l so that $\forall l_i \in l$ $l_\perp \leq_d l_i \leq_d l_\top$.

We denote level to each element l_i. To identify the level l of the dimension d we will use $d.l$. The two special levels $l_\perp$ and $l_\top$ will be called *base level* and *top level* respectively. The partial order relation in a dimension is what gives the hierarchical relation between levels.

Definition 2 *For each pair of levels l_i and l_j such that $l_j \in H_i$, we have the relation $\mu_{ij} : l_i \times l_j \rightarrow [0,1]$ and we call this the* **kinship relation***.*

If we use only the values 0 and 1 and we only allow an element to be included with degree 1 by an unique element of its parent levels, this relation represents a crisp hierarchy. If we relax these conditions and we allow to use values in the interval [0,1] without any other limitation, we have a fuzzy hierarchical relation.

Definition 3 *We say that any pair (h,α) is a* **fact** *when h is an m-tuple on the attributes domain we want to analyze, and $\alpha \in [0,1]$.*

The value α controls the influence of the fact in the analysis. The imprecision of the data is managed by assigning an α value representing this imprecision. Now we can define the structure of a fuzzy DataCube.

Definition 4 *A DataCube is a tuple $C = (D, l_b, F, A, H)$ such that $D = (d_1,...,d_n)$ is a set of dimensions, $l_b = (l_{1b},...,l_{nb})$ is a set of levels such that l_{ib} belongs to d_i, $F = R \cup \emptyset$ where R is the set of facts and $\emptyset$ is a special symbol, H is an object of type history, A is an application defined as $A : l_{1b} \times ... \times l_{nb} \rightarrow F$, giving the relation between the dimensions and the facts defined.*

For a more detailed explication of the structure and the operations over then, see (Molina et al., 2006; Delgado et al., 2004).

2.2 User View

Over the structure presented, we defined a new layer. Its main objective is to hide the complexity of the model and provide the user with a more understandable result. Using fuzzy summary operators, we define the *user view*. As an example of this type of operator, we can use the one proposed in (Blanco et al., 2003). This operator proposes the use of the fuzzy number that best fits, in the sense of fuzziness, the fuzzy set or fuzzy bag. We can use more simple operators as the *weighted average*.

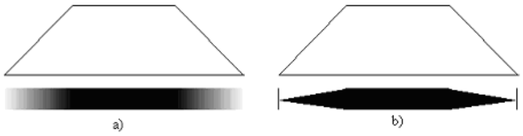

Figure 1: Graphical way to represent fuzzy numbers.

To give an intuitive way to interpret the results is important, as shown by Codd et al. in the 11th OLAP product evaluation rule ((Codd, 1993)). We propose two methods to represent fuzzy numbers in a graphical way as an user view. Both approaches are shown in Figure 1. In Figure 1.a the approach followed is to use a color gradient to represent the membership grade of the values: a clearer color means a low membership degree, and an intense color means a high membership. The other approach (Figure 1.b) consists on changing the width of a bar to represent the membership: a low membership degree is represented by a thicker bar than the one for a high degree.

2.3 *F*-Cube Factory System

In this section we comment the main characteristics of *F*-Cube Factory, the system that implements the fuzzy model proposed, in addition to others models, see (Delgado et al., 2005) for more details. The system is built using server/client architecture. The server implements the main functionality over the DataCubes (definition, management, queries, aggregation operators, user views operators, API for DataCube access, etc.). The client we have developed is web based and is thought to be light enough to be used in a personal computer and to give an intuitive access to server functionality (hiding the complexity of using a DML or DDL to the user).

The DataCube defined in next section and the examples queries over it (Section 4) have been built using this system.

3 TRADING STRATEGIES DATACUBE

In this section we present the structure of the DataCube built using the fuzzy multidimensional model presented. The DataCube built is shown in Figure 2.

3.1 Dimensions

We have defined 13 dimensions. In all of them we have used the minimum and maximum operator as t-norm and t-conorm when calculating the *extended kinship relation*. In next sections we present the structure of each one.

Strategy: We consider 107 different strategies and we classify them according to two characteristics: *Style*, been the possible values *day*, *base* and *anti*; and *Frequency*, according to if the strategy considers a *short* or *medium* period in its normal application.

Profit: This annualized performance takes into account commissions, slippages, and management expenses. The values are in the interval [-50,70]. Over these values, we have defined three categories considering if the value is bad, normal or good. There is no standard to define the edge between the labels, and most of the times experts use imprecise expressions to define then. So, we have used fuzzy concepts in the dimension to manage the relationships between the concrete values and the quality. The fuzzy intervals are represented in Figure 3. Under these circumstances, the structure of the dimension built is *Profit* = ($\{$*Values*, *Quality*,*All*$\}$,$\leq_{Profit}$,*Values*,*All*), where $\leq_{Profit}$ is the relation that defines the hierarchical relations as follows: *Values* $\leq_{Profit}$ *Values*, *Values* $\leq_{Profit}$ *Quality*, *Values* $\leq_{Profit}$ *All*, *Quality* $\leq_{Profit}$ *Quality*, *Quality* $\leq_{Profit}$ *All*, *All* $\leq_{Profit}$ *All*.

Sharpe ratio: The sharpe ratio is a measure of risk-adjusted performance of an investment asset, or a trading strategy. This variable is used to characterize how well the return of an asset compensates the investor for the risk taken. When two assets are compared, the one with the highest sharpe ratio provides a greater return for the same risk. Investors are often advised to pick investments with high sharpe ratios. This value is often used to rank the performance of portfolio or mutual fund managers. On this variable, the range of values is [-5,5].

We have defined three categories to classify according to the quality as in the previous dimension. We have considered three labels depending on the values can be considered bad, normal or good to select the trading strategy. As in other dimensions, the membership of each value to a category is not well defined, so we consider fuzzy intervals to build the kinship relations of the values (Figure 7).

The structure of the dimension is analogous to previous one: *Sharpe Ratio* = ($\{$*Values*, *Quality*,*All*$\}$,$\leq_{SR}$,*Values*,*All*).

Loss series (Drawdown): This is the greatest loss sequence, or rather, the greatest drop between the peak of accumulated profit and the lowest point. Measurement begins when the fall starts and ends when a new maximum is reached. The values of the examples are in the range [0,100] and, as can be deduced from the explanation, high values are the bad ones and the low ones are translated into a good performance of the strategy. The edges between good and normal, as well as between bad and normal, are not defined in a crisp manner. If we consider them as crisp ones, two values very near con be considered as belonging to different categories. The fuzzy intervals used are shown in Figure 4. *Drawdown* dimension is defined as follows: *Drawdown* = ($\{$*Values*, *Quality*, *All*$\}$,$\leq_{LS}$,*Values*,*All*).

Potential: This is a measure of the performance in relation to the maximum loss series, and the values belong to the interval [-2,6]. The structure of the dimension is as follows: *Potential* = ($\{$*Values*, *Quality*,*All*$\}$,$\leq_{Potencial}$,*Values*,*All*), where the kinship relations between the values of the level *Values* and *Quality* are represented in Figure 8.

Consistency: In our particular case, this variable refers to the number of negative results over time. It presents values in [-4,4]. The values below 0 and near to this value are not good because it means a large number of negative results. Values near the upper edge represent a good performance of the strategy. To characterize this behavior we define three categories: the bad values, the good ones and an interval between both than represents a normal situation. Figure 5 presents the imprecise intervals proposed. The structure of the dimension is *Consistency* = ($\{$*Values*, *Quality*,*All*$\}$,$\leq_{Consistency}$,*Values*,*All*).

Reliability: This variable represents the percentage of winning trades considering all the trades. As it is a percentage, the values are in the interval [0,100], being the greatest ones the good performance for a strategy. If the value is under the 50%, the strategy performs badly. The values in the middle are considered as normal situation (Figure 9). The structure of the dimension is analogous to previous ones.

E01 and E04: These variables are the one-year and four-year stars following the Standard & Poors' method. By dividing the strategy's average relative performance by the volatility of its relative performance, we are measuring not only its ability to outperform its peer but also to do so in a consistent way; the higher the ratio, the greater the strategy's ability to outperform its peers consistently. The number of stars depends on the relative position of the strategy according to the others considered. If a a strategy has 1 or 2 stars, it is considered a bad one; if it presents 4 or 5, it is considered a good one, and in the case of 3 stars, the strategy presents a normal behavior. In this case, the kinship relations between the values and its quality is crisp.

The structure of the dimensions are as follows: *E01* = ($\{$*Values*, *Quality*,*All*$\}$,$\leq_{E01}$,*Values*,*All*), and *E04* = ($\{$*Values*, *Quality*,*All*$\}$,$\leq_{E04}$,*Values*,*All*).

Risk: The risk combines the probability of a nega-

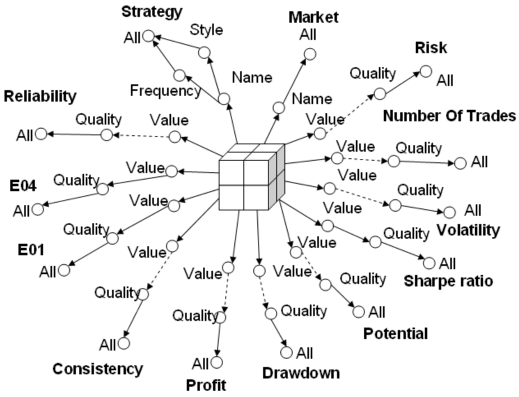

Figure 2: DataCube used in the analysis.

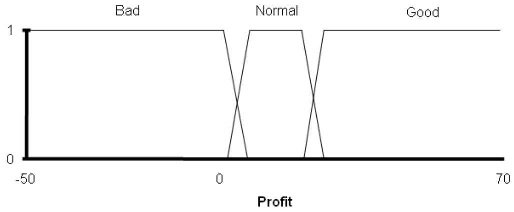

Figure 3: $\mu_{Quality,Values}$ for *Profit* dimension.

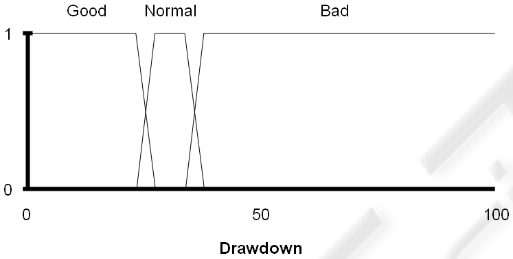

Figure 4: $\mu_{Quality,Values}$ for *Drawdown* dimension.

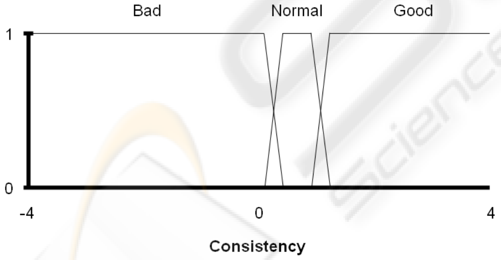

Figure 5: $\mu_{Quality,Values}$ for *Consistency* dimension.

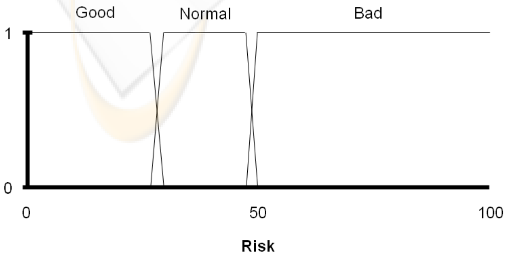

Figure 6: $\mu_{Quality,Values}$ for *Risk* dimension.

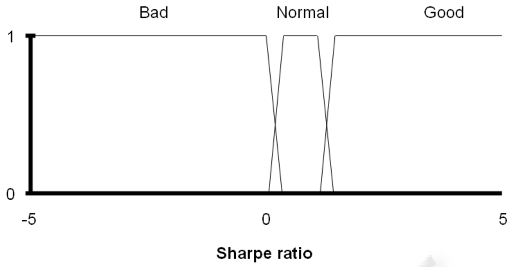

Figure 7: $\mu_{Quality,Values}$ for *Sharpe ratio* dimension.

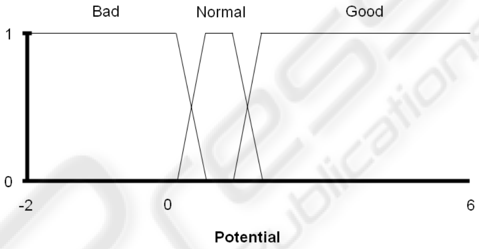

Figure 8: $\mu_{Quality,Values}$ for *Potential* dimension.

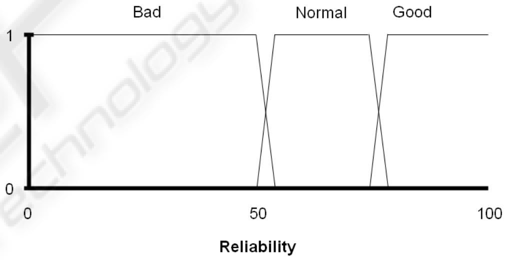

Figure 9: $\mu_{Quality,Values}$ for *Reliability* dimension.

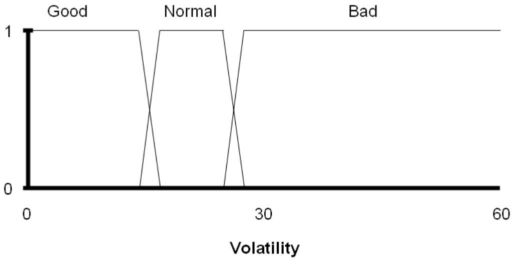

Figure 10: $\mu_{Quality,Values}$ for *Volatility* dimension.

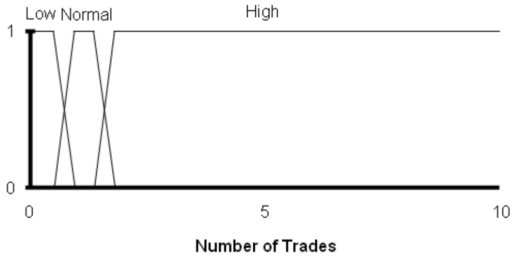

Figure 11: $\mu_{Range,Values}$ for *Number of trades* dimension.

tive event occurring with how much damage this event would case. It is measured in the interval [0,100], and the high values represents high risk. Figure 6 shows the classification of the values in bad, normal or good ones. The structure of the dimension is equal to the previous ones presented.

Volatility: Volatility is the standard deviation of the change in the value of a financial instrument with a specific time horizon. It is frequently used to quantify the risk of the instrument during this time period. Volatility is expressed in annualised terms. The values are in the interval [0,60] and we have divided it into three different categories according to the meaning of the values: good, bad and normal values. Figure 10 shows the kinship relations between the *Values* and the *Quality*.

The structure of the dimension is as follows: $Volatility = (\{Values, Quality, All\}, \leq_{Volatility}, Values, All)$.

Number of Trades (Activity): This variable models the number of trades per day. The values are in the range [0,10]. We divide the interval in three categories: low, normal or high. The two extremes are bad behavior and the center can be considered as normal, so we have a hierarchy with three levels and one fuzzy relation between the base level (*Values*) and the *Range* level. In this later case, the kinship relations are presented in Figure 11 and the structure of the dimension is as follows: $NumberOfTrades = (\{Values, Range, Quality, All\}, \leq_{NoT}, Values, All)$.

Market: The strategies can be used on different markets and may present different behavior depending on it. We have considered the strategies in the following markets: CAC-40, DAX-50, Euronext, Ibex-35 Nasdaq, Russell, name as CAC, DAX, EUR, IBX, NDQ, USA, and RUS. No hierarchy has been define over the markets, only to consider all the values together: $Market = (\{Name, All\}, \leq_{Market}, Name, All)$.

3.2 Measures

On this DataCube we have only considered one measure: the number of times the same coordinates (same value in all the base levels of the dimensions) appear in the data set.

3.3 DataCube

Finally, the structure of the DataCube is $C_{Trading} = (\{Strategy, Profit, Sharpe\ Ratio, Drawdown, Potential, Consistency, E01, E04, Risk, Volatility, Number\ of\ trades, Market\ \}, \{Name, Values, Values, Values, Values, Values, Values, Values, Values, Values, Values, Name, \}, Number \bigcup \emptyset, \Omega, A)$, where A is the relation that associates each fact with the corresponding values of the base level of the dimensions. The DataCube has been filled with 3109 facts. In next section we present some example queries over the structure and brief comments of the results.

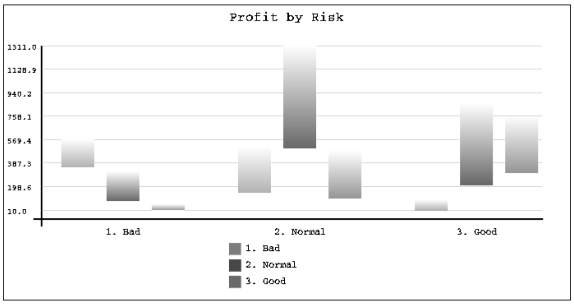

Figure 12: Profit according to the risk.

4 QUERIES

In this section we present two queries solved over the DataCube built to show the advantages of using a fuzzy model.

Query 1: We first want to know if there is a relation between the *profit* and the *risk*. The graph shown in Figure 12 represents in the coordinate axis the categories for the profit values and each one of the subcategories is the value for *risk* variable. The values in *Y* axis are the number of times each combination of values appears. According to the graph, it shows a relationship between the values of both variables: when the *profit* is *good* most of the times the *risk* is *good* too; the same as *normal* and *bad* categories. As we have defined the categories it means that for high profits the risk is low, and when the profits are low the risk is high, so there are an inverse relationship between both variables.

Speaking about the imprecision, when the *profit* has *good* values the imprecision appears in *good* and *normal* values, but very low for *bad* category, being the highest one for *normal* values of *profit*. This circumstance is due to values in the middle of both categories but nearer to the *good* one. In the case of *normal profit* we have most of the values in the *normal risk* subcategory with a higher imprecision than in the others. Under this circumstance, we can say that the values are around a *normal risk*. If we consider a crisp model the values would belong to one of the categories and we can not know the distribution of the values inside the categories, so we do not have this information for the analysis.

Query 2: The second query is intended to know the relationship between the variables *profit* and *drawdown*. Figure 13 shows a graphic representation of the facts of the resulting DataCube. In the graph the coordinate axis represents the categories for the variable *profit* and each one of the subcategories shows the number of times the strategies have this value according to the *quality* in the variable *drawdown*. This second query only present a significative relationship between the variables when the *profit* is *bad*, due to the higher the value of *drawdown* (bad values) the higher is the number of the cases where the value of *profit* is *bad*. When the *profit* belongs to category *normal* then all three categories in *drawdown* have values very similar but with different imprecision. When the *profit* is *good* it seems that values are distributed near the extreme subcategories (*bad* and *good*).

When considering the imprecision, the more significance cases are for *normal* and *good* categories in *profit*. When the values belong to *normal*, all three subcategories for *drawdown* have high imprecision. This circumstance shows that the values are distributed in the three categories, having an important number of cases with values between two categories. When the *drawdown* belongs to *normal* the imprecision is higher due to it considers values that are between this category and the other two, meanwhile the other, the imprecision is the result of the values between the own category and *normal*). When the *profit* is *good* the imprecision is centered in the *normal drawdown* due to the values are distributed mainly in the *bad* and *good* categories but with values near the *normal* one.

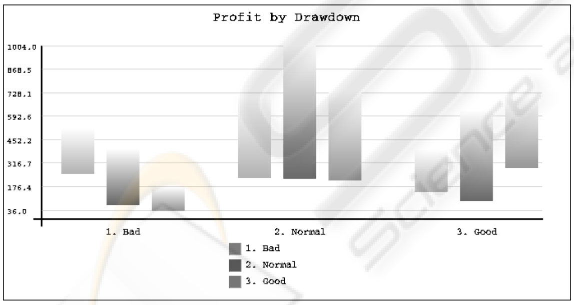

Figure 13: Profit according to the drawdown.

As a result of the analysis of this query we get that *drawdown* does not provide good information about the expected *profit*.

5 CONCLUSIONS

In this paper we have presented a fuzzy multidimensional model as an exploratory analysis tool for studying trading strategies. OLAP tools are suitable for this kind of analysis and are able to work with a high amont of data given an intuitive access for the user. Using an OLAP system that implements a fuzzy multidimensional model allows to model some concepts in a way closer to user perspective (e.g. values near the average, etc.). In this situation to have an intuitive way to show the results, that hides the complexity of fuzzy logic, is very important.

We have model an economic problem using a fuzzy multidimensional model that has enabled us to use fuzzy concepts, to obtain analysis nearer to user, and relax the edges of intervals, to reduce the edge problem. Using the *user views* the model hides the complexity of the model and gives graphic interpretation for the queries. We have obtained coherent results for the queries and have shown that, in some situation, using fuzzy hierarchies is more informative for the user, getting results that can not be obtained using a crisp model, as the distribution of the values around the edges of the categories.

REFERENCES

Blanco, I., Sánchez, D., Serrano, J. M., and Vila, M. A. (2003). A new proposal of aggregation functions: The linguistic summary. *Lecture Notes in Computer Science*, 2715:127–134.

Codd, E. (1993). Providing OLAP (On-line Analytical Processing) to user-analysts: An IT mandate. Technical report, E.F. Codd and Associates.

Delgado, M., Molina, C., Rodríguez-Ariza, L., Sánchez, D., J.M., and Vila, M. (2005). F-CubeFactory: A fuzzy OLAP system for supporting imprecision. In *IFSA 2005 World Congress: Fuzzy Logic, Soft Computing and Computational Intelligence*, volume 1, pages 635–640, Beijing (China).

Delgado, M., Molina, C., Sánchez, D., Vila, M. A., and Rodriguez-Ariza, L. (2004). A linguistic hierarchy for datacube dimensiones modelling. In *Current Ussues in Data and Knowledge Engineering*, pages 167–176, Varsovia.

Molina, C., Sánchez, D., Vila, M. A., and Rodríguez-Ariza, L. (2006). A new fuzzy multidimensional model. *IEEE Transaction on Fuzzy Systems*, 14(6):897–912.