Software Cost Estimation by Fuzzy Analogy for Web Hypermedia Applications

Ali Idri, Ph.D., ENSIAS, Rabat, Morocco

Alain Abran, Ph.D. ETS, Montreal, Canada

Azeddine Zahi, FST, Fes, Morocco

International Conference on Software Process and Product Measurements

November, 6-8, 2006, Cadiz, Spain

ENSIAS

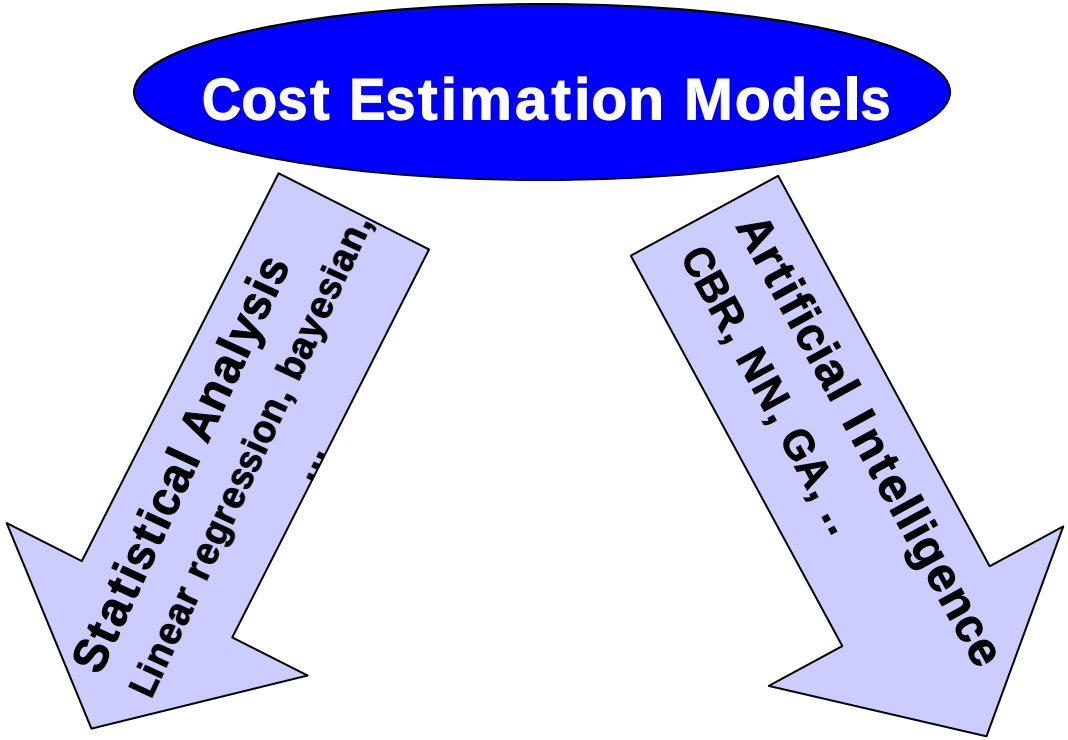
Outline

- Motivations and Objectives
- Fuzzy Analogy: An overview
- Building Fuzzy Sets for the Web Software Attributes
- Overview of Empirical Results
- Conclusions and Future Work

ENSIAS

Motivations and Objectives

- Software cost estimation is one of the most critical activities in managing software projects



Algorithmic Models **Non-Algorithmic Models**

ENSIAS

⊙**Estimation by Analogy is a promising technique to solve the software cost estimation problem:**

- **It is easy to understand and to explain its process to the users**
- **It can model a complex set of relationships between the dependent variable (cost, effort) and the independents variables (cost drivers)**

⊙ **Limitation :**

- **Estimation by Analogy cannot handle correctly the case where software projects are described by categorical data such as 'very low', 'low', 'high'....**

⊙ **Hence, we have developed a new approach: Fuzzy Analogy based on reasoning by analogy and fuzzy logic**

- **7th IEEE International Symposium on Software Metrics, London, 2001**
- **8th IEEE International Symposium on Software Metrics, Ottawa, 2002**

ENSIAS

- **We validated Fuzzy Analogy on the COCOMO'81 dataset**

- **The accuracy of Fuzzy Analogy is compared with that of three other models: classical analogy, Intermediate COCOMO'81 and fuzzy Intermediate COCOMO'81**

- **Fuzzy Analogy performs better in terms of accuracy (MMRE=21) and in its adequacy in dealing with linguistic values**

⊙ Objectives

to validate Fuzzy Analogy on a dataset containing 54 Web hypermedia applications

ENSIAS

Fuzzy Analogy: An Overview

- **Fuzzy Analogy is a fuzzification of the classical analogy procedure**
- **Fuzzy Analogy is composed of three steps:**
  - **Identification of software projects**
  - **Evaluation of similarity between projects**
  - **Adaptation**
- **Identification of software projects :**
  - **The aim is to describe the software projects by a set of attributes Vj that are measured by linguistic values** A_k^j
  - **Each linguistic value** A_k^j **is represented by a fuzzy set with a membership function** $\mu_{A_k^j}$
  - **The fuzzy sets and their membership functions are defined by using:**
    - **1) empirical techniques which construct membership functions from expert knowledge; or**
    - **2) automatic techniques, which construct membership functions from historical data using clustering techniques**

ENSIAS

⊙ Evaluation of similarity between projects

$$d(P_1,P_2)=\begin{cases} all\ of\ d_{v_j}(P_1,P_2) \\ most\ of\ d_{v_j}(P_1,P_2) \\ many\ of\ d_{v_j}(P_1,P_2) \\ at\ least\ four\ of\ d_{v_j}(P_1,P_2) \\ \ldots \\ there\ exists\ of\ d_{v_j}(P_1,P_2) \end{cases}$$

- **These measures evaluate the overall similarity $d(P_1,P_2)$ of two projects P_1 and P_2 , by combining the individual similarities of P_1 and P_2 associated with the various attributes Vj describing P_1 and P_2, $d_{v_j}(P_1,P_2)$, using Regular Increasing Monotone (RIM) quantifiers such as 'all', 'most', 'many', 'at most α', or 'there exists'**
- **The choice of the appropriate RIM linguistic quantifier, Q, depends on the characteristics and needs of each environment**
- **Q indicates the proportion of individual distances that we feel is necessary for a good evaluation of the overall distance.**

ENSIAS

Adaptation

- The objective of this step is to derive an estimate for the new project by using the known effort values of similar projects
- Two questions:
  - 1- How many similar projects will be used in the adaptation?
  - 2- How to adapt the chosen analogies in order to generate an estimate for the new project?
- In Fuzzy Analogy, we have proposed a new strategy for selecting projects to be used in the adaptation step:
  - For selecting the similar projects, we use the distances and the definition adopted in the studied environment for the proposition, ' Pi is a closely similar project to P.'
  - For the adaptation formula, the weighted mean of all known effort projects in the data set is used.

ENSIAS

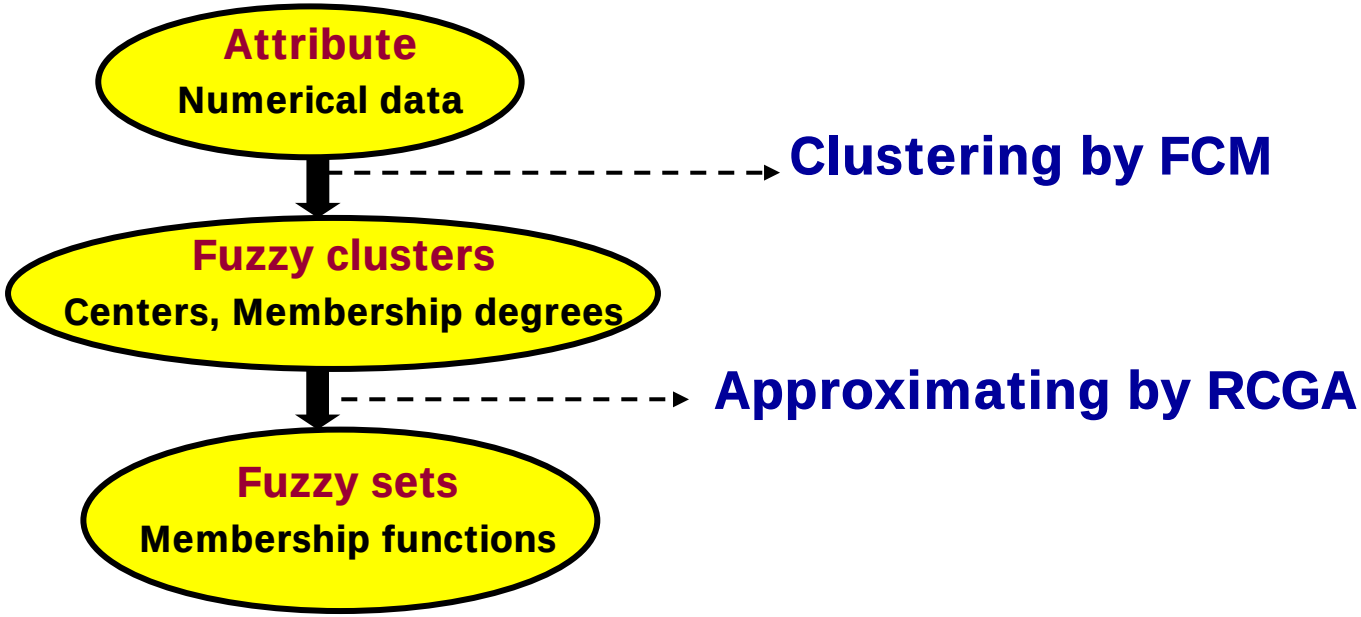
Building fuzzy sets for the Web attributes

- Each web software is described by 9 numerical attributes, such as the number of html used, the number of media files and team experience
- Initially, this dataset contains more than 9 software attributes, but some of them may be grouped together
  - Number of new Web pages developed by the team, number of Web pages provided by the customer and the number of Web pages developed by a third party (outsourced) are gouped in one attribute reflecting the number of Web pages in the application (Webpages).

Software attribute	Description
Teamexp	Average number of years' experience the team has in web development
Devteam	Number of people who have worked on the software project
Webpages	Number of Web pages in the software
TextP	Number of text pages in the software (600 words to a text page)
Imag	Number of images in the software
Anim	Number of animations in the software
Audio/video	Number of audio/video files
Tot-high	Number of high-effort features
Tot-nhigh	Number of low-effort features

ENSIAS

- The use of Fuzzy Analogy to estimate software development effort requires the determination of the fuzzy sets, and their membership functions, of the attributes describing software projects
- Because the descriptions given of the 9 Web attributes are insufficient to empirically build their fuzzy representations, we use of the Fuzzy C-Means clustering technique (FCM) and a Real Coded Genetic Algorithm (RCGA) to build the fuzzy representations of the Web software attributes
- The proposed FCM-RCGA fuzzy set generation process consists of two main steps (ICT4M, November, 21-23, Kuala Lampur, 2006)



- The FCM algorithm is a fuzzy clustering method used to generate a known number of clusters (c) from a set of numerical data

- The determination of this number is still an open problem in clustering. Often, empirical knowledge or a set of evaluation criteria is used to choose the best set of clusters. In this work, we use the Xie-Beni fuzzy cluster validity criterion proposed

- FCM is an iterative algorithm that aims to find cluster centers $(C_j), 1 \le j \le c$ and the matrix $U = (u_{ij}), 1 \le i \le n, 1 \le j \le c$ that minimize the following objective function:

$$Min\ J_m(U,C) = \sum_{i=1}^{i=n}\sum_{j=1}^{j=c}(u_{ij})^m \left\|x_i - c_j\right\|^2 \qquad \sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,\ldots,n$$

- where **m** is the control parameter of fuzziness; $U = (u_{ij})$ is the partition matrix, containing the membership values of all data in all clusters;

ENSIAS

- **After generating fuzzy sets (clusters) with their partition by means of FCM, we use an RCGA to build membership functions for these clusters;**

- **Membership functions can be trapezoidal, triangular or Gaussian.**

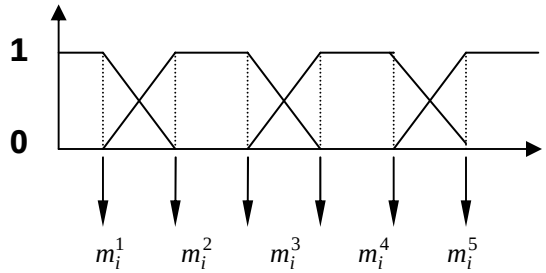
- **Our RCGA consists in building a set of membership functions that interpolates and minimizes the mean square error, which is defined as follows:**

$$MSE(\mu_1,..,\mu_c)=\frac{1}{n}\sum_{j=1}^{j=n}\left\|(\mu_1(x_j),..,\mu_c(x_j))-(u_{1j},..,u_{c,j})\right\|^2$$

- **The use of an RCGA to find membership functions requires the determination of certain parameters, such as:**
  - **the coding scheme,**
  - **the fitness function, and**
  - **the various genetic operators (selection, crossover and mutation).**

ENSIAS

- **Concerning the coding scheme, a chromosome represents the set of the unknown membership functions, , associated with the c fuzzy sets generated by the FCM**
  - **The shape of the membership functions can be trapezoidal, triangular or Gaussian.**
  - **Thus, each chromosome encodes a set of membership functions in a real vector . The genes are obtained from the shape of the membership functions.**



- **The fitness function F is obtained using the following formula:**

$$F(m_i) = \frac{MSE(m_i)}{\sum_{i=1}^{i=M} MSE(m_i)} \qquad MSE(m_i) = \frac{1}{n}\sum_{j=1}^{j=n} \left\| \mu(x_j) - y_j \right\|^2$$

- **For the three genetic operators (selection, crossover and mutation), we use those that are specific to RCGAs**

ENSIAS

Empirical Results

- The web data set contains 54 software projects
- Our similarity measures are computationally intensive; so we have developed a software prototype Matlab 7.0 under a Microsoft Windows PC environment
- This software prototype allows us to try various RIM linguistic quantifiers to the webdataset : $Q(x)=x^{\alpha}$, $\alpha>0$
- The accuracy of the estimates is evaluated by :
  - MRE

$$MMRE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Effort_{actual,i} - Effort_{estimated,i}}{Effort_{actual,i}}\right| \times 100$$

  - Pred (0.20)

$$Pred(p) = \frac{k}{N}$$

ENSIAS

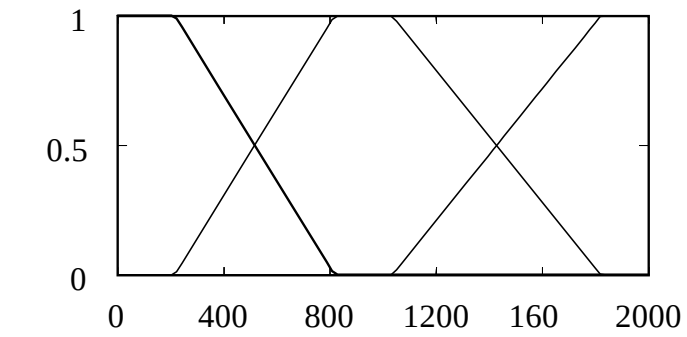
- For each Web software attribute, several experiments were conducted with the FCM algorithm, each time using a different initial matrix U. The desired number of clusters (c) is varied within the interval [2,7].

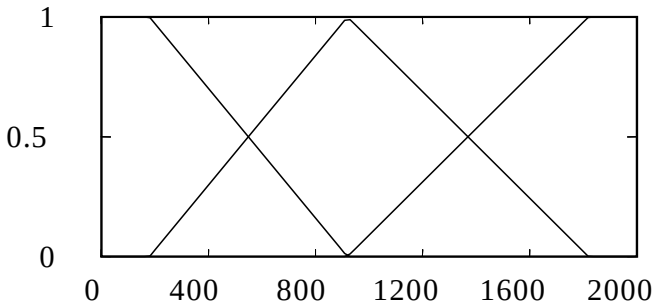
- The parameter m is fixed to 2 in all experiments.

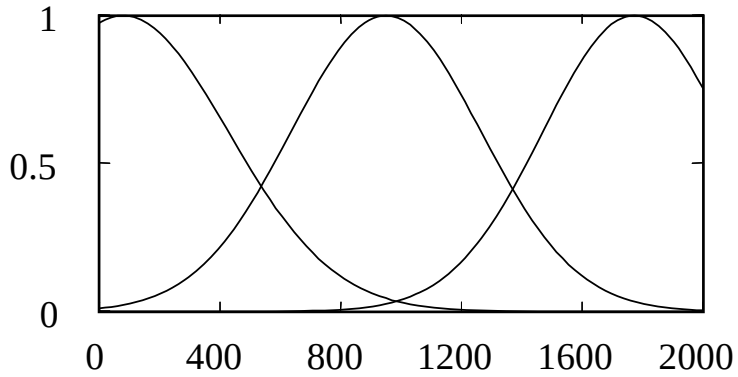
- We use the Xie-Beni criterion to decide on the number of clusters. For each attribute, we choose the number of clusters that minimizes the value of the Xie-Beni criterion.

Attributes	#fuzzy sets	Attributes	#fuzzy sets	Attributes	#fuzzy sets
DevTeam	7	TEXTP	3	Audio	4
Teamexp	7	IMAG	3	Tot-high	5
Webpages	2	ANIM	3	Tot-nhigh	5

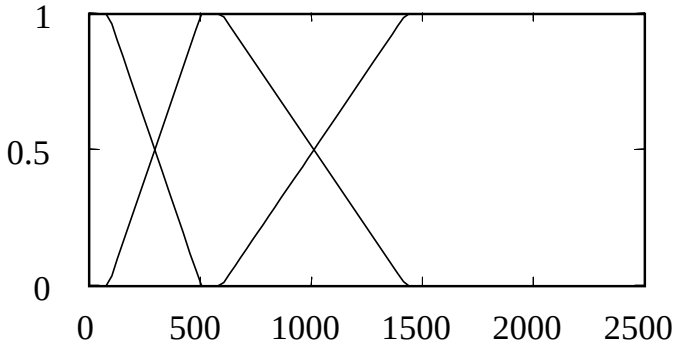
IMAG attribute

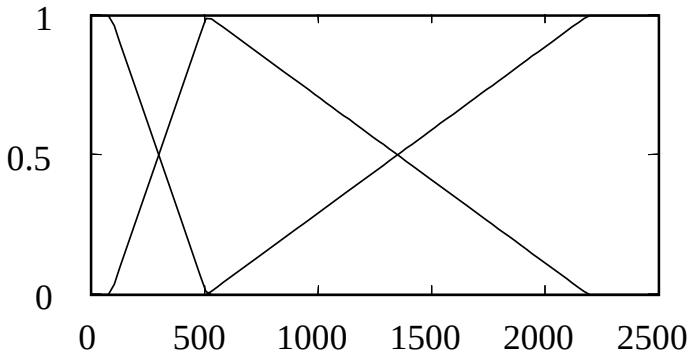


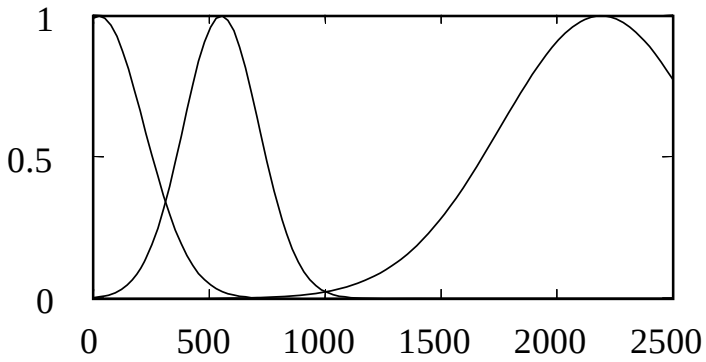




TEXTP attribute



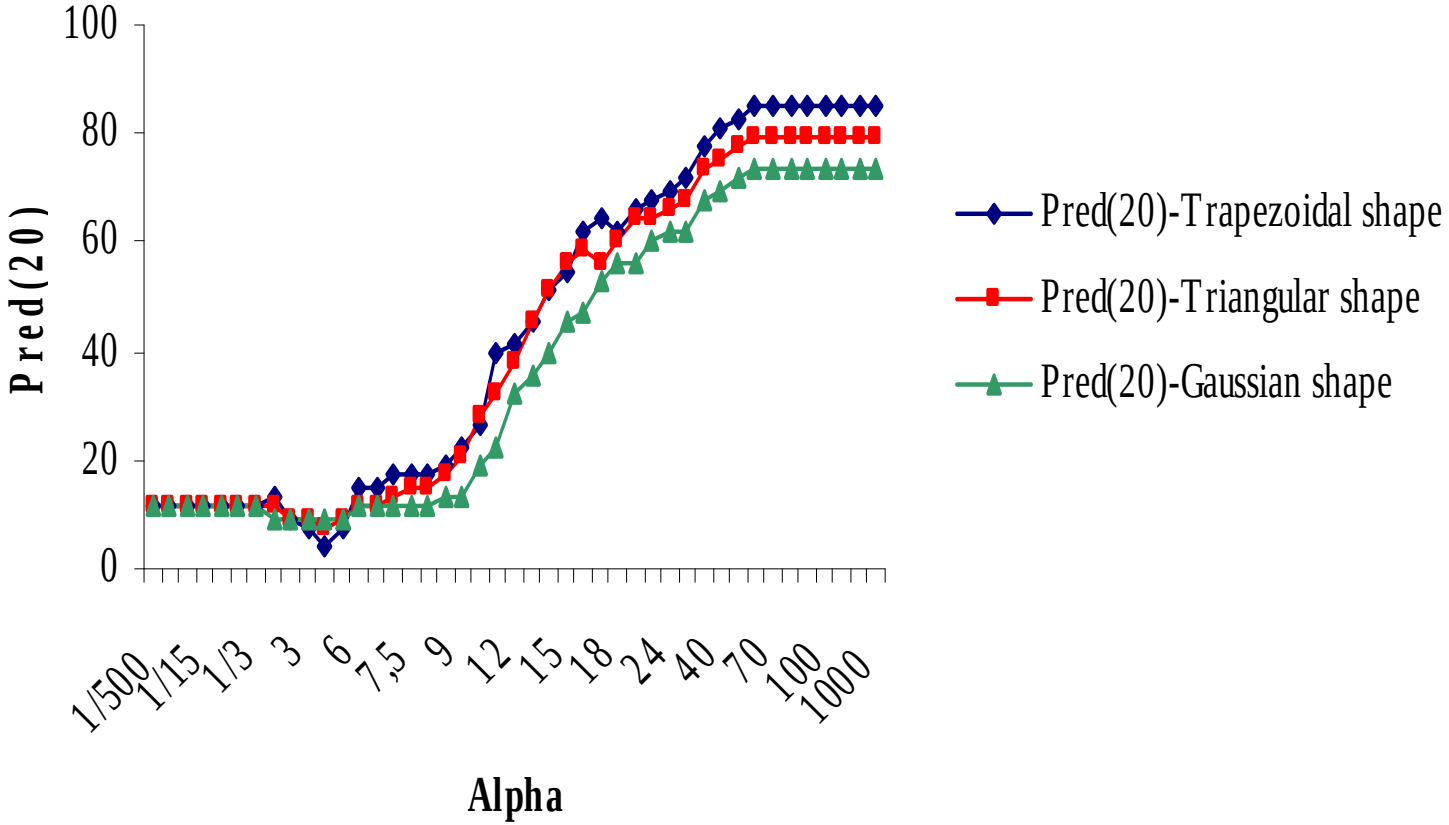




	Trapezoidal functions		Trinagulair functions		Gaussian functions	
alpha-RIM	Pred(0.20)	MMRE	Pred(0.20)	MMRE	Pred(0.20)	MMRE
1/10	11.32	717.78	11.32	718.90	11.32	719.54
1/7	11.32	711.12	11.32	712.67	11.32	713.53
1/3	11.32	683.10	11.32	686.37	11.32	688.12
1	13.21	600.72	11.32	607.91	9.43	612.33
3	7.55	434.93	9.43	445.37	9.43	455.48
7	16.98	250.73	13.21	260.14	11.32	272.85
10	26.42	177.59	28.30	184.28	18.87	195.68
15	54.72	119.31	56.60	122.88	45.28	131.48
25	71.70	78.36	67.92	81.12	62.26	87.93
30	77.36	69.91	73.58	73.17	67.92	79.75
40	81.13	62.19	75.47	65.75	69.81	72.40
50	83.02	59.71	77.36	63.30	71.70	70.17
60	84.91	58.92	79.25	62.54	73.58	69.46
70	84.91	58.68	79.25	62.30	73.58	69.28
80	84.91	58.60	79.25	62.23	73.58	69.22

- **The accuracy of the estimates depends on the linguistic quantifiers (α) used in the evaluation of the similarity between projects**
- **When α tends towards zero => the overall similarity takes into account fewer attributes among all describing software projects.**
- **When α tends towards infinity => the overall similarity takes into account many attributes among all describing software projects**

ENSIAS


100
80
60
40
20
0
Pred(20)
1/500
1/15
1/3
3
6
7,5
9
12
15
18
24
40
70
100
1000
Alpha
Pred(20)-Trapezoidal shape
Pred(20)-Triangular shape
Pred(20)-Gaussian shape

ENSIAS

Conclusion and Future Work

- We have validated the Fuzzy Analogy approach for estimating the cost of Web hypermedia applications
- We have used the Fuzzy C-Means clustering technique (FCM) and a Real Coded Genetic Algorithm (RCGA) (the FCM-RCGA process) to build the fuzzy representations for the Web software attributes
- The membership functions generated may be trapezoidal, triangular or Gaussian.
- The results of this validation show that Fuzzy Analogy generates accurate estimates, using trapezoidal, triangular or Gaussian fuzzy representation.
- We are currently looking at comparing Fuzzy Analogy and classical analogy on the Web dataset
- Another interesting avenue of research would be to look at the accuracy of Fuzzy Analogy when using the FCM-RCGA process rather than empirical knowledge for building fuzzy sets **(SETIT, March, 2007)**.

ENSIAS

Thank you

ENSIAS