Archaeological Applications of Fuzzy Databases

Franco Niccolucci[1] – Andrea D'Andrea[2] – Marco Crescioli[3]

Contents

Abstract

This paper deals with problems concerning statistical data (e.g. deriving from archaeometry) in an archaeological database, when an unaware use may lead to erroneous conclusions. A new model is proposed for these cases, using fuzzy logic to assign a reliability coefficient to imprecise attributes. Considering a case study, we generalize the assignment of age, gender and chronology to burials. The procedures are general and can be fruitfully used also in other investigations. To manage these fuzzy attributes, we personalized a free Relational Database Management Systems (RDBMS) and created a WWW interface to ease data consultation and allow remote access.

1. Quantitative applications and archaeological theory

In a recent paper (Barceló 2000), J. Barceló wisely pointed out that the applications of computers to archaeology have arrived at an elevate level of complexity, often characterized by sophisticated and expensive techniques, but such resources are still not fully exploited for their investigation potential, notwithstanding the goals achieved especially in spatial technologies and virtual reality applications. For the Spanish scholar, the low use of these advanced computer technologies in archaeological research derives from the fact that we are not able to ask questions complex enough for so complex instruments and therefore archaeological results still lack.

Pursuing the application in archaeology of the most recent hardware solutions and of the most promising software developments, produced by research or by the market, often generates technical systems, which are efficient and reliable but are not accompanied by an adequate level of theoretical and methodological reflection.

Behind a shiny technological apparatus it is often hidden a preoccupying trivialisation, caused by the absence of reflection on the impact of the use of advanced technology on the process of historical knowledge. However, critical elements pervade the archaeological use of virtual reality and emerge also towards inter-site GIS systems, oriented only to environmental variables and therefore deterministically biased. Stating the importance of the connection between the improvement of computer applications and archaeological research, Harris and Lock pointed out that a GIS system is

[1] University of Florence, Florence, Italy – e-mail: niccolucci@unifi.it
[2] CISA – Istituto Universitario Orientale, Naples, Italy – e-mail: dandrea@iuo.it
[3] Unirel srl – Sesto Fiorentino, Italy – e-mail: marco@unirel.it

not impartial or neutral: it " *represents the social reproduction of knowledge and, as such, the development of a GIS methodology cannot be divorced from the development of the theory needed to sustain it*" (Harris and Lock 1995:355).

Very recently, a similar attention to a correct definition of the correlation between techniques and interpretative processes seems to characterize also mathematical and statistical applications. These have been for a long time the main quantitative application in archaeology and now they seem to undergo a new growth, after a decline due to the crisis of the processual approach, which had represented their theoretical and methodological basis (Moscati 1996).

The recent contribution of quantitative methods (Buck, Cavanagh and Litton 1996; Delicado 1999) such as non-parametric statistics, Bayesian statistics and fuzzy theory, certainly helped to invert the negative trend that had characterized quantitative archaeology around the middle of the eighties under the post-processual criticism (see, for instance, Hodder 1982). Perhaps the negative reaction to the use of statistics to support interpretation may have generated a new relationship between archaeology and mathematics.

A new quantitative approach is based on evaluating the impact of statistical-mathematical models on carrying on archaeological research (Moscati 1996; Voorrips 1996; Wilcock 1999), not only as far as data analysis and classification is concerned, but also in formalizing procedures and in the use of statistical sampling techniques. Thus, the post-processualist image of the computer as a neutral instrument sides with the New Archaeology vision of it as an objective meter of historical and human facts and behaviours: both these approaches, only apparently opposed, in fact converge to the same negation of the importance of computers in archaeological theory and method. Taking in no consideration the deep connection between computational methods and their impact on archaeological theory leads inevitably to a cul-de-sac: the blind pursuit of the "discovery" of "innovation" and novelty without understanding the function of the proposed solution in the process of historical and archaeological investigation, and the consequent inability of going beyond the proposal of toys, which so often are as expensive as useless.

Hopefully a different approach, as the one we suggest for fuzzy theory, may represent a useful step to envisage a new and more promising relationship between archaeological theory and practice and the use of models deriving form other disciplines (Crescioli, D'Andrea, Niccolucci *in press*). In our opinion, it is not correct to choose a quantitative technique only because it seems to fit better with the current investigation, since this attitude produces inevitably a mechanic, and unreflective, application of quantitative techniques that may lead to erroneous conclusions. By choosing a technique, we must bear in mind that in this way we are making a cognitive choice that will reflect on data and results. Fuzzy theory reminds us continuously that during an investigation we make choices that are determinant to formalize data but leave no sign in the interpretative process, so that raw data and hypothetical or reconstructed information become unscindible: the more the formalism used for data analysis is hidden, as in computer applications and, in particular, in database applications, the bigger it is the risk of overwhelming the original information content of data with the subjective meaning of interpretation.

2. Databases and archaeological theory

The huge amount of data that characterize any archaeological investigation and the pervasive presence of computers in every aspect of present life have ultimately led to a generalized use of DBMS's (Data Base Management Systems) to manage excavation data as well as any other kind of archaeological records. It appears nowadays quite natural to store and search archaeological information into a computer memory, due to the highly structured nature of forms used to record them, a condition that perhaps precedes the advent of computers but certainly is enforced by their use. These tools undoubtedly have a great importance in easing archaeological data management and the synthesis process, so that nowadays even the most conservative educational institutions cannot any more exclude some database training from archaeologists' curricula. Using DBMS has thus become a part of current archaeological practice and little attention is therefore paid to its implication on the correctness of data. Sometimes this is due to an excessive confidence in automatic processing, sometimes it is the ignorance of simple statistical laws about error propagation that may induce to false conclusions, that moreover have the aspect of indisputable truth since they have been produced by a machine, which by assumption makes no mistake in computations. After they have been recorded into a database, archaeological records loose any element of uncertainty and subjectivity and become as trustworthy as the computer itself.

This consideration should not imply a luddist rejection of computers, which by the way are not guilty of the wrong results, but simply the awareness that computations on uncertain data follow rules that differ from ordinary ones, with or without a computer; even the simple act of counting is no more the same. In other words, since archaeological data have an intrinsic uncertainty, any conclusion drawn basing on them cannot ignore elementary statistical rules, including the paradox that 1 + 1 not always makes 2.

In fact, every time you recognize something there is some uncertainty in the attribution and if you repeat this process several times, as it happens for instance when classifying archaeological finds, errors associated to each item add up, giving a total error that in some cases may be unacceptable.

In most cases one can ignore this feature, because the error is so small that deterministic rules and statistical rules in practice do not differ: however, this should not be given for granted in every case.

This reasoning has particular implications when using a DBMS to record data. Usually this is accomplished by crossing boxes, or filling fields according to standardized dictionaries, and there is no space for uncertainty or doubts. One has to decide to cross the box labelled "black" or the one labelled "white", with no possibility of grey. Then *alea iacta est*, the die is cast, and that choice will obliterate forever the real archaeological record and will be processed with many similar ones, possibly thousands of them as it happens when managing finds from an excavation. The computer, in its cold assurance, will keep no track of the archaeologist's human hesitation.

Thus the subjective attribution is unconsciously objectivized and different levels of reliability are equalized to absolute certainty by the magic of computers. Should we not introduce an alert that some of the data are "more subjective" than others and even the archaeologist who originally interpreted them, trusted them at different degrees? Probably yes, and it is a common practice to mark by interrogation marks less reliable attributions. But interrogation marks are difficult to process, and in no way supported by DBMS. So our proposal aims at introducing some attributes that make evident the reliability of data, and a few simple and transparent rules to process them.

Ignoring the problem of data reliability is still worse when they are derived from statistical processing. This happens when archaeology uses the results of other scientific techniques, as in archaeometry, and our case study will illustrate one such example.

In conclusion, databases are very useful in recording archaeological data and using them in every day's archaeological practice is an achievement that should be not discussed. But a naive use may lead, in a few cases, to incorrect conclusion, that can be prevented with some simple technical improvement. Our contribution hopefully moves towards this perspective, by simply quantifying (in an absolutely subjective way) how much the compiler of the database trusted the data, and consequently giving some reasonable rules to process this reliability coefficient through all the computations the database is used for. It must be pointed out that the numeric nature of this reliability coefficient should in no way be interpreted as an "objective" measure of the uncertainty but only an expression of the compiler's reliability subjective evaluation. Therefore, it should be clearly stated in the accompanying documentation the meaning of different numeric values, and how they are computed when the coefficient derives from computation, as it will happen in our case study.

Together with other practices, as the generalized disclosure of archaeological databases to the public, this approach will moreover contribute to guarantee the correctness of application of the scientific method, which requires the possibility of tracking back, at least in theory, the inference of results from data, beyond the "black box" of the database.

3. The case study

The present paper considers the data resulting from a sample of burials discovered in the cemetery of Pontecagnano, an important Etruscan-Campanian settlement placed about 70 Km South of Naples. The funerary area, extending below the modern centre, in over than forty years' investigations produced more than 100 burial nucleuses and more than 8000 tombs dating between the First Iron Age (9th century BC) and the Hellenistic period (beginning of the 3rd century BC). To manage the huge quantity of finds, the archaeological team is carrying on a GIS project since some years (D'Andrea 1999). The project consists of a cartographic database, implemented with Mapinfo, whose main function is to place exactly the ancient remains on modern cartography and to store topologic, spatial and alphanumeric data of each tomb and burial area.

The burials examined in the present paper pertain to the most recent phases of use of the Etruscan-Campanian cemetery. They were edited by Serritella (Serritella 1995) in a volume which includes the philological study of grave-goods, the analysis of most significant pottery production and, above all, the reconstruction of the ancient community of Pontecagnano in the 4th and 3rd century BC starting from the analysis of funerary customs. The tombs studied by Serritella are distinct from the remainder of the cemetery and insist upon free surfaces, not occupied in previous periods, thus constituting a privileged observatory for the study of the society of the Hellenistic period. Of all the tombs, 65% had grave-goods in them, while the remaining 35% did not, including about 7% which had been certainly violated in ancient time.

In order to examine the funerary behaviours, the author uses the analysis of pottery and burial typology and, moreover, the results obtained determining the gender and age of the deceased by classical anthropometric methods (Scarsini and Bigazzi 1995; Petrone 1995). These are based on statistical values that may be obtained with different procedures, giving as results different numeric coefficients. In particular, the gender coefficients vary within a range from +2 and –2: positive values refer to male gender, the negative ones to females. Unfortunately, most of the values do not reach these extremes: only 11 are outside (–1, +1), that is about 20% of the cases for which an osteological coefficient can be evaluated and 13% of all cases, so for most cases the level of uncertainty is rather high.

Notwithstanding the uncertainty of the palaeo-anthropological results, obtained with a statistical computation applied to the dimensions of each skeleton, the tables that compare grave-goods, gender and age, to reconstruct the horizontal stratigraphy (age classes) and the vertical stratigraphy (social status) of burial areas, do not show the variability of anthropological determinations. So the statistical information turns into certain data.

Correctly, the publication gives all the details of the anthropological analysis so that the reader may check the scientific results, but all this is irreparably lost when data are stored in a database.

To circumvent this drawback, our proposal suggests to use the statistical information already available to keep the coefficient variability within data structure, by creating special attributes and showing how to process them.

To this aim, we have analyzed the frequency distribution of the osteological coefficients obtained in the case study, dividing [-2, 2] into intervals of length 0.1. It should be expected to obtain a bi-modal distribution, with peaks corresponding to the two most frequent values denoting males and females.

The histogram of this frequency distribution is shown in figure 1.

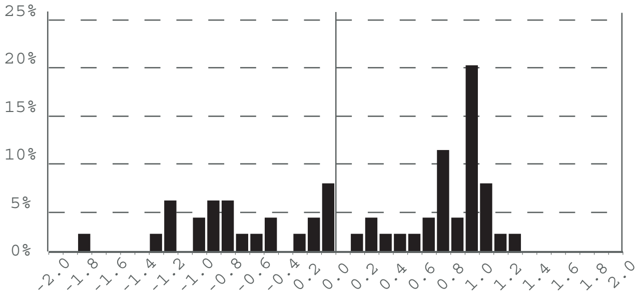


Figure 1 – Frequencies of gender coefficients

As it can be easily verified from figure 1, the distribution of gender coefficients is only roughly bimodal: the male modal value is +1, while female coefficients have no mode. Moreover, even adding the frequencies of modal values and the nearest neighbours, the total does not reach 30%.

Probably such characteristics may be influenced by the choice of the interval width: using 0.2 instead of 0.1 gives in fact a better double-bell-shaped curve (with still low frequencies of the modal values) but it confirms that discriminating the gender by means of osteological coeffcients is not a straightforward task .

4. Fuzzy set concepts

We are not going to deal in detail here with fuzzy theory, referring for further details to Crescioli, D'Andrea and Niccolucci (*in press*) and the bibliography included. It will suffice to remind that, given a set *A*, a **fuzzy entity** is the couple formed by a variable *X* having values *x* in *A* and a function f_X from *A* to [0, 1], so that to any instance *x* of *X* it is associated a number $f_X\ (x)$ in [0, 1], which can be interpreted as the degree of reliability of *x*, and will therefore be named in the sequel the **(fuzzy) reliability coefficient** attached to *x* while *f* will be named the **(fuzzy) reliability function**. So, a fuzzy entity extends the concept of variable by adding these reliability coefficients.

In particular, a **fuzzy label** is such a couple, the first one assuming nominal values (the labels). For instance, fuzzy gender is a fuzzy label, with nominal values "male" and "female", each one having a number attached, the fuzzy reliability coefficient of the assignment.

A **fuzzy value** is another kind of fuzzy entity, in which the first element of the couple, the variable, has a numeric range. Fuzzy age is such, being formed by a possible range of ages, each one having a corresponding fuzzy reliability coefficient.

Fuzzy labels can be fully described as arrays, having the labels in the first column and the corresponding reliability coefficients in the second. Fuzzy values can be represented in the same way if the range of possible values is finite; otherwise a function from *A* to [0, 1] is needed. A typical form of the function is trapezoidal as shown in figure 2.

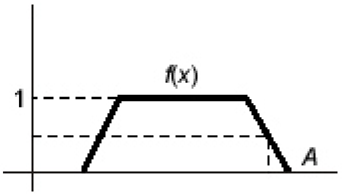


Figure 2. Graph of a trapezoidal fuzzy reliability function *f*

Also the concept of equality needs an extension to operate with fuzzy entities.

We first introduce the *similarity* $s(x)$ between the fuzzy entities X and Y, respectively with fuzzy reliability functions f, g, defined over the same domain A, which is the (numerical) function

$$s: A \to [0, 1]: s(x) = \min (f(x), g(x)), x \in A.$$

A graphical representation of $s(x)$ is the following, in which it is assumed that X and Y are fuzzy values so that A is a numerical set, and both f and g have a very simple, trapezoidal form shown in figure 3.

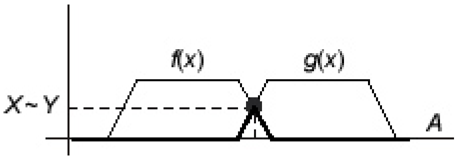


Figure 3. Graph of similarity function s (heavy line) and value of $X \sim Y$ (marked point)

To compare globally the two fuzzy entities X and Y, the maximum of s over A is taken: in this way, we define a **fuzzy operator**, that is a function associating a number in [0, 1] to each couple of fuzzy entities. We shall use the symbol ~ to denote this operator, called *fuzzy equal* to. In the previous picture, the value of $X \sim Y$ is given by the ordinate of the marked point in figure3.

The rationale of this definition depends on the interpretation of the fuzzy reliability function: taking the minimum of the two functions means that for each possible value in A we consider the worst condition for each fuzzy entity, but globally speaking, the most likely situation corresponds to the greatest of these values. The equality of the two items may derive from being both "male" or, independently, both "female": so the reliability of the equality, regardless of what case determines it, is larger than the reliability of each single case, which adds confidence to the overall reliability. A prudential approach will give for the reliability of each case the minimum of the two reliability coefficients, and the overall reliability will equal the greatest of the two, with no additional contribution from the other. For another example, consider two disjoint age intervals X and Y: strictly speaking, there is no equality between them, since the parts in which they have a reliability of 1 are disjoint, but both have overlapping tails in which they are less likely, however not impossible: the common value where they have the highest likelihood is the marked point of figure 3.

The definition of fuzzy equality is an example of generalization to fuzzy entities of familiar concepts as equality, counting, adding, averaging, and so on. We are not going to deal with these concepts any more: only counting fuzzy quantities will be taken into account in this paper. To count occurrences, that is to compute frequencies, we need to generalize the familiar operation of counting, that is adding one when the desired result comes out (for instance, "female" when counting gender occurrences) and adding zero, instead, when it does not (that is, the result is "male"). In our generalized model, we shall total the fuzzy coefficients for each case, so that the count of each possible outcome will be the sum of the fuzzy coefficients. This agrees with common sense weighting in average evaluation and is also a particular case of a more general "Extension principle" (see Yager and Filev 1994:16-18).

5. Fuzzy entities in the case study

In the case study, three attributes have been recognized as fuzzy entities: gender and age of the deceased, buried in the tomb, and the chronology of the burial.

Gender may be considered as a fuzzy label, as stated before, while age and chronology are fuzzy values. For each one of these fuzzy entities we shall briefly explain how to evaluate the second member of the couple, the reliability coefficient: for fuzzy gender, this will imply the evaluation of two numbers, one for each gender, based on the osteological coefficient, while the other two attributes require the definition of a function, as shown below.

There are several (in fact, infinite) possible ways to assign numerical values to the gender coefficients; the one we choose are based on the following considerations:

- In this case study, few osteological coefficients (less than 20%) go beyond +1 or –1, which can be considered the best possible results in these conditions.
- When the male coefficient gets its highest value, the female one should get the lowest, and vice versa.
- When the osteological coefficient varies between –1 and +1, the corresponding fuzzy gender coefficient increases (or decreases) uniformly.

So, denoting by *k* the osteological coefficient and by *m* and *f* the male and female corresponding gender coefficients, to derive *m* and *f* from *k* we can build the function shown in table 1 or, graphically, in figure 4.

k	m	f
(–2, –1)	0	1
(–1, +1)	0.5*k* + 0.5	–0.5*k* + 0.5
(+1, +2)	1	0
Undefined	0.5	0.5

Table 1. Derivation of *m* and *f* (reliability gender coefficients) from *k* (osteological coefficient).

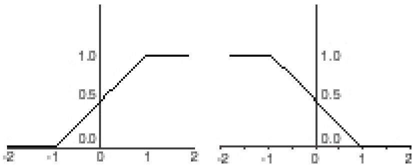


Figure 4. Derivation of *m* (left, ordinates) and *f* (right, ordinates) from *k* (abscissae)

Thus we are able to obtain the fuzzy gender coefficients for each item from the value of its osteological coefficient. The resulting fuzzy gender attribute will be an array as already noted, for instance {(male, 0.8), (female, 0.3)}.

Notice that, even if there is, in general, no mutual dependence between *m* and *f*, the definition we chose implies that *m* + *f* = 1.

There are some cases in which the value of *k* is undefined since there were no elements enough to apply the osteological method. In these cases, our choice is to assign a value of 0.5 to each of the gender coefficients, that is *m* = *f* = 0.5. This assignment is based on the fact that the gender of the deceased is undecidable on the known elements, so that both gender are equally likely (or unlikely). The disadvantage of this choice is that the difference, if any, between the case of *k* = 0 and *k* not computable is lost, so somebody could prefer a different assignment as, for instance, *m* = *f* = 0. We really see no relevant loss on information and the method applies with both choices, so the one we adopted will have no consequence on the model validity.

The osteological determination of age ranges is based on two different methods, producing in several cases conflicting results, as in the case of tomb 4046 (Scarsini and Bigazzi 1995:139, Tab.1a) for which the two estimates are respectively of 50 ± 2 years and 20 – 25 years.

While the results of the second method are given in the form of a range with no other information, for the first one the authors publish (Scarsini and Bigazzi 1995:139-140) the central value μ and the standard deviation σ, so it is reasonable to assume that estimated ages have a Gaussian distribution with mean μ and standard deviation σ as given in the paper. Since the area below the Gaussian curve between $\mu - \sigma$ and $\mu + \sigma$ equals 68% of the total area (see any text on statistics, for instance Mood, Graybill and Boes 1979), we may conclude that the estimated age values between $\mu - \sigma$ and $\mu + \sigma$ are the most probable and hence the most reliable, with a tail, on both sides, having lower probability. In terms of reliability of the result, we may therefore assume that this is the highest for the values within $[\mu - \sigma , \mu + \sigma]$, decreasing to zero outside; to keep things simple, the usual trapezoidal shape may be used, so that the reliability coefficient will be 1 within $[\mu - \sigma , \mu + \sigma]$, going to 0 at $\mu - \sigma - \theta$ and at $\mu + \sigma + \theta$, θ being a positive number. In order to estimate θ, the table of the normal distribution tells us that when $\theta = 0.5\sigma$, 7% of the area is left out on each side; for $\theta = 0.64\sigma$, the remainder is 5%; for $\theta = 1.33\sigma$ it is 1% and so on, the last two values being those normally used for confidence intervals in hypothesis testing. The choice among these possibilities is subjective, and prudentially we went for $\theta = 0.5\sigma$, considering unreliable those cases that have a probability less than 0.07. Since the age range has a width of 2σ, this means that in our trapezoidal approximation, we allow for a slack of 25% of the length of the estimated age interval, on each side of it, assigning a fuzzy function with value 1 on the interval determined by osteology, which descends to 0 on both sides with constant slope. The same rule will be applied to the second osteological method.

For instance, an osteological estimate of the age range as 20 – 40 will correspond to the fuzzy age represented by the function shown in figure 5 and explained in table 2.

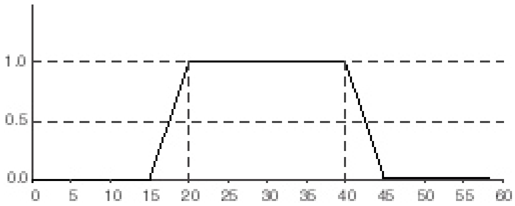


Figure 5. Fuzzy age coefficient (example for the age range 20 – 40)

Age x	Fuzzy coefficient
< 15	0
$15 \leq x < 20$	$0.2x - 3$
$20 \leq x < 40$	1
$40 \leq x < 45$	$-0.2x + 9$
$45 \leq x$	0

Table 2. Relationship between the fuzzy age coefficient and age (example for the age range 20 – 40)

The general formula for f in terms of k, μ and σ can be easily computed from the above rule and results as shown in figure 6:

$$f(k)=\begin{cases}0 & \text{per } k \le \mu - 1.5\sigma \\ 1-\dfrac{(\mu-\sigma)-k}{\sigma/2} & \text{per } \mu - 1.5\sigma < k \le \mu - \sigma \\ 1 & \text{per } \mu \text{ - } \sigma < k \le \mu + \sigma \\ 1-\dfrac{k-(\mu+\sigma)}{\sigma/2} & \text{per } \mu + \sigma < k \le \mu + 1.5\sigma \\ 0 & \text{per } \mu + 1.5\sigma < k\end{cases}$$

Figure 6 Fuzzy coefficient *f* as a function of the mean μ and the standard deviation σ of the estimated osteological age

When the osteological investigation gives two distinct, not overlapping ranges for the age, they will correspond to a fuzzy reliability function built considering the two separate parts as distinct reliability functions *f*(*x*) and *g*(*x*) and defining their fuzzy OR as follows:

f OR *g* (*x*) = max (*f*(*x*),*g*(*x*))

with the graph shown in figure 7.

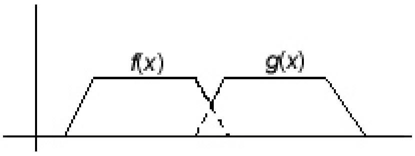


Figure 7. Graph of fuzzy OR (heavy line)

This set of rules implies that to wider age intervals it corresponds a greater uncertainty and consequently a larger slack and has other consequences that are worth considering.

Let us consider two cases, the first one with an osteological age range of 22 – 30 years and the second with a range of 23 – 39: which one should be considered "younger"? Intuition says the first one, but this is not true. Since the ranges have a statistical nature, there are tails for both, that our cautious assumption determines in 2 years for the first one and 4 years for the second one (traditional statistical assumptions would have fixed them in 2.56 years and respectively 5.12 years, at a confidence level of 5%, and even larger for a confidence level of 1%). This implies that the complete age range (with tails) is 20 – 32 for the first case and 19 – 43 for the second: who is "younger", now?

This simple example shows that our common sense reasoning may be fallacious and new categories need to be introduced even for simple comparisons, which loose any significance when applied to statistical data.

As far as chronology is concerned, in this context nominal constants had been used, as usual, for instance "First quarter of 4th century BC", meaning the time interval [–400, –375].

Each of these has therefore been converted into a fuzzy value, with a fuzzy coefficient given in the usual trapezoidal for each time range, that is equal to 1 for the corresponding time interval and a tail δ on each side, on which the fuzzy reliability coefficient varies uniformly from 0 to 1 or vice versa.

Again, several choices are possible for δ and candidates are 6.25 years (25% of the range as for age), 8 years (32% of the range, corresponding to a confidence statistical level of 0.1) 16.625 years (66.5% of the range, corresponding to a confidence statistical level of 0.01). We choose 15, the rounded value next to the latter, to express the high degree of indeterminacy, in numerical terms, of the chronological traditional assignment, which leaves more space for tail values outside of the interval. this is equivalent to assume a slack of 60% of the time range length. Figure 8 shows an example.

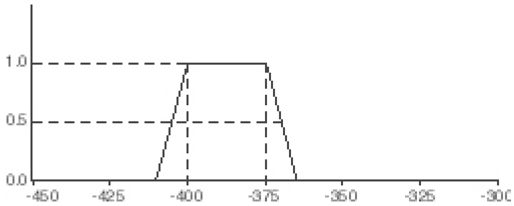


Figure 8 Fuzzy chronology with a slack of 15 years (example).

It might be argued that wider intervals give less confidence to each single value of a fuzzy value: an age interval as 20 – 40 would make less credible every single age, 35 for instance. If this is the case, confirmed by the osteological method used to determine the age interval, the top value of the reliability function should be lowered accordingly. But with no other information, all values should have equal, and top, reliability, as we suggest.

This argument possibly derives from a misunderstanding between the probability of a single value, which is lower if more cases are possible, and the fuzzy reliability of each single case, which is not influenced by the number of them since there is no constraint of adding up to 1. This is the reason why we use this approach and not a probabilistic one: from a probabilistic point of view, the probability of an exact value is zero, as it is well known, even if an exact value, perhaps unknown to us, should have existed. So in this context archaeology does not need to deal with low probability figures, but with "exact" figures having a different reliability, to establish archaeological inference knowing how reliable they are.

Concluding this paragraph on the evaluation of the fuzzy reliability coefficients, we want to underline that the subjective character of their choice has ultimately emerged. In our opinion, however, this is their strength and not their weakness: it has been definitely proved that any attempt to construct a completely deterministic model, which should give archaeological results only by means of computations, is fallacious. Subjective, in this meaning, differs from arbitrary, and is strictly related to the concept as used in De Finetti (1970) or in Savage (1972), who prefers the term personalistic to denote this approach.

6. Operating in the database with fuzzy entities

To store the data in a database, it is necessary to create special data types corresponding to fuzzy entities, that is fuzzy labels and fuzzy values, and to give rules to process them. Since these have already been defined in Crescioli, D'Andrea and Niccolucci (*in press*), we refer to that paper for a more detailed description.

As noted before, a fuzzy label is an array of couples formed by a label, that is a nominal element, and a number in [0, 1]. The nominal elements are chosen from the common domain, so different instances of the fuzzy label consist of the same nominal elements with possibly different coefficients. If we agree to put always the nominal elements in the same (not relevant) order in the array, a fuzzy label is characterized by the domain, that is the common set of the *n* possible labels, and a *n*-ple of numbers, differently valued for each instance of the label. In our case, the fuzzy gender is therefore characterized by the domain, the two labels "male" and "female", and a couple of numbers as (0.3, 0.7), having conventionally agreed that the first on refers to "male" and the second one to "female".

The definition of each member of the data type `FUZZY_LABEL`, as `FUZZY_GENDER`, requires therefore to store somewhere the domain, that is {"male", "female"} in this case, and then consists of a one-dimensional array of real numbers.

For fuzzy values, the necessary function is approximated by a piecewise linear function, so that only the corner points need to be stored. Previous models used only "trapezoidal" functions as the ones shown in the above figures, but this limits the field of application, as shown in Crescioli, D'Andrea and Niccolucci (*in press*): our models puts no such restriction.

So, the FUZZY_VALUE data type consists of a two-dimensional array of real numbers. For example, the age interval 20 – 40 is represented as the array {(15, 0), (20, 1), (40, 1), (45, 0)} according to the assignment of the fuzzy age function stated in paragraph 5; these values are, in fact, the coordinates of the corner points of the graph of the reliability function.

It may be useful to define constants for any type of fuzzy data, which are stored in a separate table. Due to SQL naming rules, and to make easier to use constants in queries, they are denoted as functions with no parameter, for instance YOUNG().

To determine the values of these fuzzy constants, we refer to commonly used ranges by anthropologists, so that, for instance, YOUNG() means an age included within 15 and 20 years, with a slack of 2 years before and after. Naturally, constants may be modified or be defined with other values, so long as their value is clearly stated.

Finally, we need to define the operator fuzzy equal, denoted with ~. This follows the definition given in paragraph 4, and the result is a number in [0, 1]. Therefore, the comparison between two homogeneous fuzzy entities or a fuzzy entity and a constant will produce a list of numbers, corresponding to the values of fuzzy equality for different instances. For instance, the comparison between the attribute FUZZY_AGE and the fuzzy constant YOUNG() will give a list of numbers, each representing the similarity of the fuzzy age of each record to the constant (fuzzy) value chosen for YOUNG(). A dummy result for the query FUZZY_AGE ~ YOUNG() is represented in table 3, where we put the descriptions instead of the values of fuzzy entities to ease readability.

Age interval (osteological)	Possible age range (fuzzy)	Young constant	Result of FUZZY_AGE ~ YOUNG()
10 – 18	8 – 20	15 – 20	1
22 – 30	20 – 32		0.5
40 – 48	38 – 50		0
22 – 38	18 – 42		0.666
22 – 22	22		0
…			…

Table 3. Results of FUZZY_AGE ~ YOUNG() (example)

The apparently counter-intuitive result that 22 – 30 is "less similar" to YOUNG than 22 – 38 is a consequence of the statistical nature of age ranges and is perfectly coherent with the fuzzy treatment of data, as already noticed in paragraph 4.

Fuzzy counting, as defined in paragraph 4, does not require any special function, it simply uses SUM.

7. Implementing the fuzzy database

Implementing the fuzzy database requires an extensible DBMS. We chose for this PostgreSQL, a RDBMS available under Linux operating system, since it is fully relational, it is free software and is customisable, in the sense that new data types, functions and operators can be added to the standard ones.

PostgreSQL can be queried within a terminal window, using `psql`, a command line SQL environment with standard features. In our case we used `psql` to create the new data types, to define the database structure and to load the data, which were available and had previously been typed, verified and converted to text format. Any software can be used for this, and we did not develop a graphical interface because we did not need it to input the data, as direct conversion was quicker. Data were then manipulated to give expressions as the following:

```
INSERT INTO TOMB VALUES(85, 4012, 'Maisto', 'Cappuccina', 'FALSE', 'TILES', 40, 216, 70, 'SE-NW',
'FALSE', 'M','', '{0.9, 0.1}', 'A', '46-52; 40-45', '{{38.75, 0}, {40, 1}, {45, 1}, {45.45, 0.64},
{46, 1}, {52, 1}, {53.5, 0}}', '1st quarter 3rd cent. BC', '{{-285, 0}, {-300, 1}, {-275, 1}, {-260,
0}}', 'TRUE', 'SUPINE', 'INHUMATION');
```

The above SQL expression assigns values to all the fields of the record, most of which are not fuzzy and have not been dealt with in the present paper. `TOMB` is the name given to the table, values in bold refer to fuzzy attributes and italics to the corresponding "original" expression, which are kept into the database for comparison. In this case, the (osteological) gender was M for "male"; the (osteological) age, according to the two different methods, consisted of the two distinct, not overlapping intervals 46 – 52 and 40 – 45 (a contradiction if manipulated with traditional methods and also impossible to manage with previous fuzzy database models), and the chronology was rendered as explained above. Creating table `TOMB` required the previous definition of the fuzzy data types, which was accomplished thanks to the `psql` command `CREATE TYPE` that allows the definition of personalized data types.

In a similar way the table `GRAVEGOODS` was created with data concerning grave-goods.

The constants are inserted into the table `CONSTANTS` and then they are used to create functions and operators, as fuzzy equality $\sim$. This operator is based on the function `f_equal(x,y)`, the only piece of software written in C to make computation quicker, and it is computed according to the definition given in paragraph 4. It has been introduced in order to allow an expression of the form `f_age ~ ADULT()`, which is much easier to understand than the equivalent (and cumbersome) expression `f_equal(f_age,'{{16,0},{21,1},{40,1},{45,0}')`.

8. The web interface to the database

Apart from data input, every search on the database may be performed using a web interface, which accesses the database locally or remotely, via a local network or the Internet.

By the way, a similar interface could be developed for any database (regardless of the presence of fuzzy attributes) and would make working with the data easier and quicker. Since it is based on a powerful RDBMS as PostgreSQL, it has also many processing advantages on commercial programs as Access or Filemaker.

The installation requires a Linux-powered server; however, since no graphical interface is required on the server, almost any PC may do the job, even an older model which otherwise would be substituted with a newer one for everyday work and then, perhaps, stay abandoned in the attic to cover with dust: for instance, we tested an old laptop with a Pentium 166Mhz and 16M RAM and it worked flawlessly.

Using the web interface, connection to the database requires only to use of a web browser as Netscape or Internet Explorer and the most common operations, as simple queries or browsing the archive using pre-prepared forms (see figure 9, 10 and 11) that make the job quick and easy. For the future, more web pages are planned to accomplish other search functions.

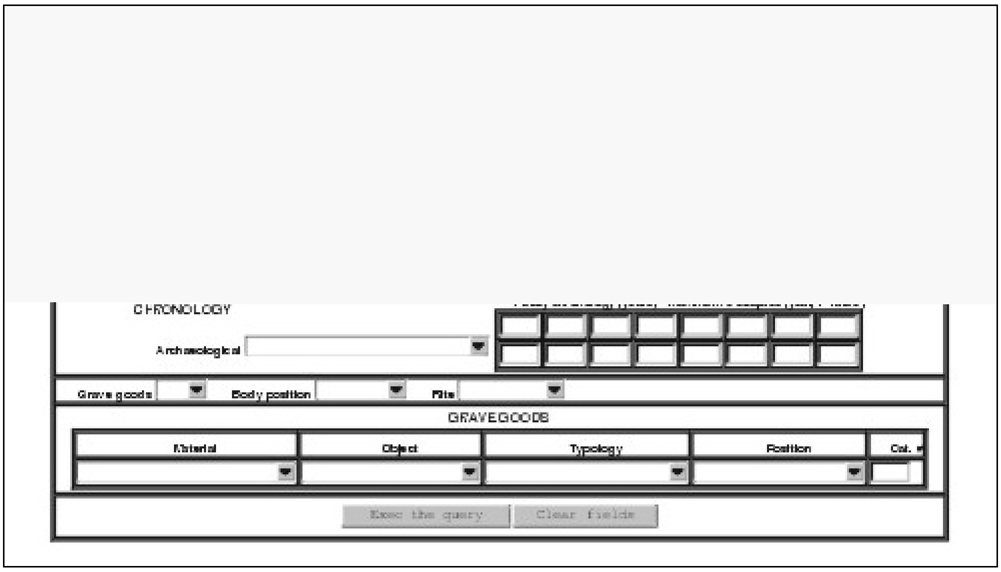
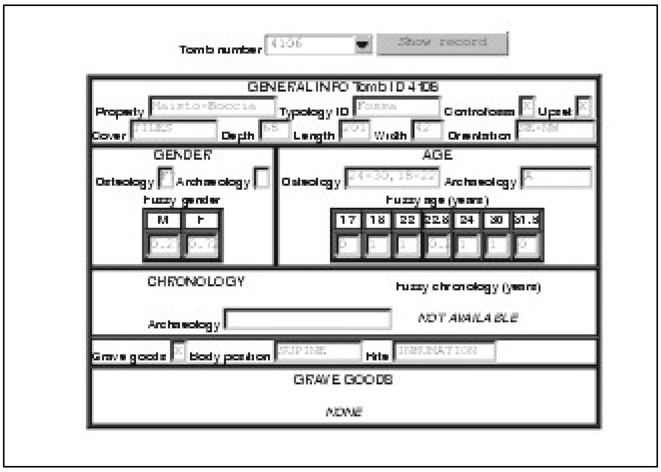


Figure 9. Form used to set query conditions, with some fields selected for display and one condition set on property.

Figure 10. Result of previous query. For space reasons, the list has been cut. Tomb numbers link to the full record.



Tomb number	4106

GENERAL INFO Tomb ID 4106

Property	Maisto-Boccia
Typology ID	Fossa
Controforma	X
Upset	X
Cover	TILES
Depth	65
Length	201
Width	42
Orientation	SE-NW

GENDER

Osteology	F
Archaeology	

Fuzzy gender

M	F
0.2	0.7

AGE

Osteology	24-30, 18-22
Archaeology	A

Fuzzy age (years)

17	18	22	22.8	24	30	31.5
0	1	1	0.2	1	1	0

CHRONOLOGY

Fuzzy chronology (years)

Archaeology

NOT AVAILABLE

Grave goods	X
Body position	SUPINE
Rite	INHUMATION

GRAVE GOODS

NONE

Figure 11. Tomb record, accessed from the previous list or directly selecting the number.

The forms have been written in PHP3, another freely available software module that has a nice interface with PostgreSQL and generates HTML pages that can be saved by the user with the "Save as" function of the browser for off-line use. This is another strong point for the choice of this RDBMS.

The database containing the case study data can be actually accessed via Internet using the web interface. Since this web publication is experimental and aimed only at an illustration of the present paper, the actual URL may vary in the future, so interested visitors are advised to ask the authors for the current address of the page, from where it is possible to download the fuzzy functions as well.

9. Archaeological application to the case study

It might be reasonable to ask, at this point, what is the impact on archaeological research of all the machinery set up in previous paragraphs. Even if the present paper aims at contributing at a methodological level, we present in this last paragraph some evidence of the results that can be obtained using fuzzy models.

Among the information obtained from the "archaeology of the death", data on demography are traditionally considered as "natural" or biological: However, as d'Agostino (1985:52) noted, also anthropological information must be interpreted within the archaeological framework, since the definition of age classes and male or female roles, only apparently objective, must be always referred to the social context in which it originated. Thus, demographic data cannot be mechanically taken into account, basing only on sociological assumptions. On the contrary, they should always be compared with information derived from the analysis of funerary customs, in order to avoid to overimpose categories deduced from the "community of the alive" on the "world of the dead".

Within demographic applications to cemetery interpretation, two criteria are particularly significant to verify how much the funerary sample is representative (d'Agostino 1990).

The first one is based on the ratio between the number of adults and the number of children and on the ratio between the number of males and the number of females: each of these ratios, for pre-industrial societies, should be approximately equal to 1 (Weiss 1973). When one of the sampled values is substantially different from this model, it may be concluded that the funerary sample is not representative of all the components existing in the community: in this case the sample of tombs may reflect the adoption of discriminating burial practices.

Another important criterion for cemetery analysis is based on funerary variability (O'Shea 1984). It is based on the principle that if social statuses of the deceased are present uniformly, this means that the tomb sample represents only one social class of the community. Interpreting funerary variability presents more complex problems than the above ratio criterion: indeed, hierarchy may be more or less emphasized according to the economic and social structure of the community and becomes inadequate when egalitarian ideologies prevail.

In the past, applications of statistical methods to the study of funerary custom has been based on a merely quantitative logic. Mathematical models have been used to measure "objectively" the richness of a tomb, to determine the funerary variability or to identify the social hierarchy on the basis of "energy expenditure" (on the subject see Cuozzo 1994:268; Cuozzo 1996). We suggest to use, instead, fuzzy logic to estimate how much a funerary sample is representative of the community using palaeo-anthropological demographic data. A few examples are given below, and we consider this perspective a potentially substantial contribution to carry on an analysis aiming at determining horizontal and vertical stratigraphy. Since the age and gender coefficient result from statistical computations on osteological parameters, as shown above, an "improper" use of them, or the mechanical assumption that they represent an "objective" truth, may lead to an unconscious variation of the real ratios children to adults or males to females, thus turning awry any deduction derived from the sample.

Consider table 4, derived from those published in Serritella (1995:116, 121, 123), counting the occurrences of age categories, shown as percentage of the total. The tombs are grouped by modern land owner, which gives a rough indication of their space position in the cemetery. In other words, land property, represented in the database with the name of the modern owner, is a simple and approximate, but effective, clustering of the tombs. It has been shown by Serritella that this grouping reflects significant differences in chronology or rite.

Property	Infant	Children Young	Adult Elderly	Indeterminate	Total	Ratio A/(I+C)
Maisto-Boccia	22.9%	0.0%	54.3%	22.8%	100.0%	2.37
Rossomando	33.3%	6.7%	53.3%	6.7%	100.0%	1.33
Tascone-Di Dato	16.0%	0.0%	76.0%	8.0%	100.0%	2.75
Total	22.7%	1.3%	61.3%	14.7%	100.0%	2.56

Table 4. Age classes by land property (percentages). Derived from Serritella (1995:116, 121, 123)

Indeterminate gender assignments add up to 15% of the total and thus they may significantly change the confidence of the sample: for the second group, assigning all the indeterminate cases to infants or children gives a ratio of 1.19, turning this sample into a very representative one.

Using fuzzy coefficients, we have to introduce the fuzzy age class "Infant or Child", ranging from 0 to 20 years, and the class "Adult or Elderly", ranging from 21 upwards, and compare the data with these two new constants.

Then we compare Serritella's results with shown in table 5 and easily obtainable by means of an SQL query on the database.

property	Children	Adult	Ratio A/C
Maisto-Boccia	43.9%	56.1%	1.28
Rossomando	45.3%	54.7%	1.21
Tascone-Di Dato	28.5%	71.5%	2.50
Total	41.0%	59.0%	1.44

Table 5. Age classes by land property (percentages) and their ratio, as computed from the fuzzy database

The "Rossomando" and "Maisto-Boccia" groups fit with the model, while "Tascone-Di Dato" does not. The latter shows a prevalence of adult burials.

Considering the male to female ratio, from Serritella's work we obtain the values shown in table 6
and from our database the ones shown in table 7.

Property	Male	Female	Indeterm.	Ratio M/F
Maisto-Boccia	47.4%	42.1%	10.5%	1.13
Rossomando	75.0%	25.0%	0.0%	3.00
Tascone-Di Dato	63.1%	31.6%	5.3%	2.00
Total	58.7%	34.8%	6.5%	1.69

Table 6. Gender by land property (percentages). Derived from Serritella (1995:116, 121, 123)

Property	Male	Female	Ratio M/F
Maisto-Boccia	53.1%	46.9%	1.13
Rossomando	70.4%	29.6%	2.38
Tascone-Di Dato	57.6%	42.4%	1.36
Total	57.2%	42.8%	1.34

Table 7. Gender by land property (percentages) as computed from the fuzzy database

From this table, the “Maisto-Boccia” group shows as a representative sample, “Tascone-Di Dato” is less representative and “Rossomando” is not.

Combining the two tables, only “Maisto-Boccia” remains as a representative sample, while the others present some discrimination: “Tascone-Di Dato” for age, “Rossomando” for gender. The explanation of this discrimination in terms of the social status and age and gender roles can derive only from the investigation of the grave-goods, to understand the underlying cultural model of social representation, but this goes beyond the aim of the present paper. Comparing our results with Serritella’s, what she suggested is confirmed by us, with a much higher level of confidence due to the absence of indeterminate cases that in her tables should suspend every conclusion.

10. Conclusions

If the model we propose here will be accepted, at least in the negative, if not in the positive, that is provoking greater caution when using statistical data in archaeological investigations, giving always for granted the reliability of archaeometric data will no more be possible, at least when these are critical for the validity of the interpretation model. We hope, however, that our model, together with the computer tools we made available (possibly improved by future work) will help research. Computer friendliness and the large availability of software tools had the positive effect of spreading their application but allowed, at the same time, their unaware use. Maybe this contribution will increase the archaeologists’ awareness that even when using databases, the quickest solution is rarely the cleanest one.

References

BARCELÓ, J. A., 2000. Visualizing what might be: an introduction to virtual reality techniques in archaeology. In J. A. Barceló, M. Forte, D. H. Sanders (eds.) *Virtual Reality in Archaeology*, 9-36. Oxford: Archaeopress (BAR International Series 843).

BUCK, C. E., CAVANAGH, W. G. and LITTON, C.D., 1996. *Bayesian Approach to Interpreting Archaeological Data*. New York: Wiley (Statistics in Practice).

CUOZZO, M., 1994. *Patterns of organisation and funerary customs in the cemetery of Pontecagnano (Salerno) during the orientalising period*. Journal of European Archaeology 2, 2:263-298.

CUOZZO, M., 1996. *Prospettive teoriche e metodologiche nell'interpretazione delle necropoli: la post-processual archaeology*. AION ArchStAnt n.s., 3:1-39.

CRESCIOLI, M., D'ANDREA, A. and NICCOLUCCI, F., *in press.* A GIS-based analysis of the etruscan cemetery of Pontecagnano using fuzzy logic, in Lock G.R. (ed.), *Beyond the Map: Archaeology and Spatial Technologies*, European University Centre for Cultural Heritage, Ravello, Italy, October 1-2 1999. Amsterdam: IOS Press.

D'AGOSTINO, B., 1985. *Società dei vivi, comunità dei morti: un rapporto difficile*, DialArch 1.3, III s.:47-58.

D'AGOSTINO, B. 1990. Problemi di interpretazione delle necropoli, in R. Francovich and D. Manacorda (eds.), *Lo scavo archeologico dalla diagnosi all'edizione, III ciclo di lezioni sulla Ricerca applicata in Archeologia* Certosa di Pontignano (Siena), 6-18 novembre 1989, 401-420. Firenze: All'insegna del giglio.

D'ANDREA, A., 1999. *Il GIS nella produzione delle carte dell'impatto archeologico: l'esempio di Pontecagnano*. Archeologia e Calcolatori 10:227-237.

DELICADO, P., 1999. Statistics in Archaeology: New Directions. In J. A. Barcelò, I. Briz and A. Vila (eds.), *New Techniques for Old Time, Proceedings of the CAA98 Conference*, 29-37. Oxford: Archaeopress (BAR International Series 757).

DE FINETTI, B., 1970. *Teoria delle Probabilità. Sintesi introduttiva con appendice critica*, Torino: Einaudi. English translation: *Probability thory: A Critical Introductory Treatment*, New York: Wiley, 1974.

HARRIS, T. M. and LOCK , G. R., 1995. Toward an evaluation of GIS in European archaeology. The past, present and future of the theory and applications. In G. Lock and Z. Stancic (eds*.), Archaeological and Geographical Information Systems: a European Perspective*, 349-365. London: Taylor & Francis.

HODDER, J., 1982. *Symbols in action*. Cambridge: Cambridge University Press.

MOOD, A. M., GRAYBILL, F. A. and BOES, D. C., 1979. *Introduction to the Theory of Statistics*, New York: McGraw-Hill International Edition.

MOSCATI, P., 1996. *Archeologia Quantitativa: Nascita, sviluppo e "crisi".* Archeologia e Calcolatori 7:579-590.

O'SHEA, J., 1984. *Mortuary variability*. New York: Academic Press.

PETRONE, P. P., 1995. Analisi paleodemografica e paleopatologica delle tombe in proprietà Rossomando. In Serritella 1995, Appendix I:129-134.

SAVAGE, L. J., 1972. *The Foundation of Statistics.* New York: Dover (second edition).

SCARSINI, C. and BIGAZZI, R., 1995. Studio antropologico dei resti umani. In Serritella 1995, Appendix II:135-148.

SERRITELLA, A., 1995. *Pontecagnano. II.3. Le nuove aree di necropoli del IV e III sec. a. C.*, Annali del Dipartimento di studi del Mondo Classico e del Mediterraneo antico dell'Istituto Universitario Orientale, Quaderno n. 9, Napoli.

VOORRIPS, A., 1996. *Information Science in Archaeology: a Short History and Some Recent Trends.* Archeologia e Calcolatori 7:303-312.

YAGER, R. R. and FILEV, D. P., 1994. *Essentials of Fuzzy Modeling and Control* J. Wiley & Sons, New York.

WEISS, K. M. 1973. *Demographic Models for Anthropology*, Washington: Memoirs of the Society for American Archaeology 27.

WILCOCK, J. D.,1999. Getting the Best Fit? 25 Years of Statistical Techniques in Archaeology. In L. Dingwall, S. Exon, V. Gaffney, S. Laflin, M. Van Leusen (eds.), *Computer Applications and Quantitative Methods in Archaeology 1997*, 19-27. Oxford: Archaeopress (BAR International Series 750).