Archaeological Applications of Fuzzy Databases

Franco Niccolucci

University of Florence

Via Pisana 56, 50143 Firenze, Italy

e-mail: niccolucci@unifi.it

Andrea D'Andrea

CISA – Istituto Universitario Orientale

V.tto I, S. Maria ad Agnone 8, 80139 Napoli, Italy

e-mail: dandrea@iuo.it

Marco Crescioli

Unirel SRL

Via Volturno 12, 50019 Sesto Fiorentino (FI), Italy

e-mail: marco@unirel.it

Abstract

This paper deals with problems concerning statistical data (e.g. deriving from archaeometry) in an archaeological database, when some unsuspecting application 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.

Key words: fuzzy set, fuzzy database, excavation database, necropolis, Etruscan

1. Quantitative applications and archaeological theory

In a recent paper (Barceló 2000), J. Barceló wisely pointed out that computer applications in archaeology have reached an elevated level of complexity, often characterized by sophisticated and expensive techniques; and yet such resources are still not fully exploited for their investigative potential, notwithstanding the goals achieved especially in spatial technologies and virtual reality applications. For the Spanish scholar, the sparse use of these advanced computer technologies in archaeological research derives from the fact that we are not able to ask questions complex enough for such complex instruments. Consequently, archaeological results are still lacking.

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

A preoccupying trivialisation is often hidden behind a shiny technological apparatus; it is 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 intersite 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 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, similar attention to a correct definition of the correlation between techniques and interpretative processes seems to characterize also mathematical and statistical applications. For a long time these have represented the main quantitative application in archaeology; and now, after a decline resulting from the crisis of the processual approach, which had represented their theoretical and methodological basis (Moscati 1996), they seem to be undergoing a new growth.

The recent contribution of quantitative methods (Buck et al. 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 1980s under the influence of 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 while carrying out archaeological research (Moscati 1996, Voorrips 1996, Wilcock 1999), not only as far as data analysis and classification are 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 opposing, in

fact converge to the same negation of the importance of computers in archaeological theory and method. Regarding 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 to bypass the proposal of toys, which so often are as expensive as they are useless.

Hopefully, a different approach from the one we suggest for fuzzy theory may represent a useful step for foreseeing a new and more promising relationship between archaeological theory and practice, as well as the use of models deriving from other disciplines (Crescioli et al. 2000). In our opinion, it is not correct to choose a quantitative technique only because it seems to correspond better with the current investigation; this attitude inevitably produces a mechanic and unreflective application of quantitative techniques that may lead to erroneous conclusions. By choosing a technique, we must bear in mind that we are thus making a cognitive choice, which will reflect on data and results. Fuzzy theory continuously reminds us that during an investigation we make choices that are determinant to formalize data, but they leave no sign in the interpretative process, so that raw data and hypothetical or reconstructed information become inseparable: the more the formalism used for data analysis is hidden, as in computer applications and, in particular, in database applications, the bigger 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) for managing excavation data, as well as any other kind of archaeological records. Nowadays, it appears quite natural to store and search for archaeological data in the memory of a computer, 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 serve a great purpose in easing archaeological data management and the synthesis process, so that nowadays even the most conservative educational institutions can no longer exclude database training from archaeologists' curricula. Using DBMS has thus become a part of the current archaeological practice and little attention is therefore paid to its implications on the correctness of data. On occasion, this is due to excessive confidence in automatic processing, while sometimes it is the ignorance of simple statistical laws concerning error propagation that may induce false conclusions; moreover, these very conclusions have the aspect of indisputable truth, since a machine, which by assumption makes no mistake in computations, produces them. After they have been recorded into a database, archaeological records lose any element of uncertainty and subjectivity and become as trustworthy as the computer itself.

This consideration should not imply a luddite rejection of computers, which by the way are not guilty of such erroneous results, but simply the awareness that computations based on uncertain data follow rules that differ from ordinary ones, with or without a computer. Even the simple act of counting is no longer the same. In other words, since archaeological data have an intrinsic uncertainty, any conclusion drawn on their foundation cannot ignore elementary statistical rules, including the paradox that 1 + 1 does not always make 2.

In fact, every time you recognize something there is some uncertainty in the attribution. And in repeating this process several times, as occurs for instance when classifying archaeological finds, errors associated to each item add up, producing a total error that may be unacceptable in certain cases.

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 taken for granted in every case.

This reasoning has particular implications when using a DBMS to record data. Usually, checking boxes or filling fields according to standardized dictionaries accomplishes this, and no space remains 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 forever obliterate 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 objectified and disparate levels of reliability are equalized to absolute certainty by the magic of computers. Should we not introduce a warning that some of the data are "more subjective" than others, including even the archaeologist who originally interpreted them and trusted them at different degrees? Probably yes. And it is a common practice to mark less reliable attributions with question marks. But interrogation marks are difficult to process, and in no way are they supported by DBMS. So our proposal aims at introducing some attributes that make the reliability of data evident, as well as 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; our case study will illustrate one such example.

In conclusion, databases are very useful for recording archaeological data, and using them in everyday archaeological practice is an achievement that need be not discussed. But a naive usage may lead, in certain circumstances, to incorrect conclusions; this can be prevented with some simple technical improvements. Our contribution hopefully moves towards this perspective by simply quantifying (in an absolute subjective way) how much the compiler of the database trusted the data, and consequently, by giving some reasonable rules to process this reliability coefficient through all the computations for which the database is used. 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, rather only as an expression of the compiler's reliability regarding subjective evaluation. Therefore, the meaning of different numeric values should be clearly stated in the accompanying documentation, as well as how they are computed when the coefficient derives from computation, as it will happen in our case study.

This approach, together with other practices, such as the generalized disclosure of archaeological databases to the public, will further contribute towards guaranteeing the correctness of application of the scientific method, which necessitates the possibility of backtracking, at least in theory, and 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-Campagnian settlement situated about 70 km south of Naples. The funerary area, extending below the modern town centre, produced in over forty years' investigations 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 enormous quantity of finds, the archaeological team is carrying on a GIS project lasting already a few years (D'Andrea 1999). The project consists of a cartographic database, implemented with Mapinfo, whose main function is to exactly position the ancient remains on modern cartography and to store topologic, spatial and alphanumeric data regarding each tomb and burial area.

The burials examined in the present paper pertain to the most recent phases of use of the Etruscan-Campagnian cemetery. Serritella (Serritella 1995) edited these burials in a volume, which includes a philological study of the grave goods, an analysis of the most significant pottery production and, above all, a reconstruction of the ancient community of Pontecagnano during the 4th and 3rd centuries BC, starting with an analysis of funerary customs. The tombs studied by Serritella are distinct from the remainder of the cemetery and are situated in free areas that were not occupied in previous periods, thus constituting a privileged observatory for studying the society of the Hellenistic period. Of all the tombs, 65 % revealed grave goods from within, while the remaining 35 % did not, including about 7 % which had certainly been violated already in ancient times.

In order to examine funerary behaviours, the author uses the analysis of pottery and burial typology, as well as the results obtained - concerning 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 using different procedures, resulting with a variety of numeric coefficients. In particular, the gender coefficients vary within a range of +2 and –2: positive values refer to male gender, the negative ones to female. 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 in 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, so as 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.

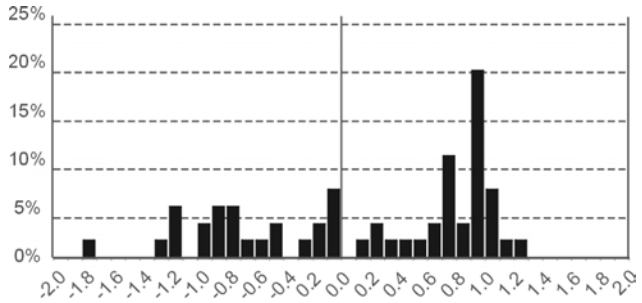

Figure 1: Frequencies of gender coefficients.

Correctly, the publication provides 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 the use of statistical information already available, by creating special attributes and showing how to process them, so as to keep the coefficient variability within the data structure.

This being our goal, we have analysed the frequency distribution of the osteological coefficients obtained in the case study, dividing [-2, 2] into intervals of length 0.1. A bi-modal distribution, with peaks corresponding to the two most frequent values denoting males and females, is anticipated. The histogram of this frequency distribution is shown in figure 1.

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

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

4. Fuzzy set concepts

We shall not go into detail here regarding fuzzy theory; further details are attainable in Crescioli et al. (2000) and the bibliography included. Let it suffice 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]. Hence, to any instance *x* of *X*, a number $f_X(x)$ in [0, 1] is associated, which can be interpreted as the degree of reliability of *x*, and will henceforth be named the *(fuzzy) reliability coefficient* attached to *x*, while *f* will be named the *(fuzzy) reliability function*. So, a fuzzy entity extends the concept of any 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 the nominal values "male" and "female"; each one has a number attached, which represents 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.

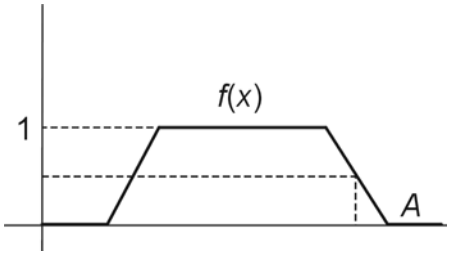

Figure 2: Graph of a trapezoidal fuzzy reliability function f.

Fuzzy labels can be fully described as arrays, where the labels are set 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 this function is trapezoidal, as shown in figure 2.

The concept of equality also 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

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

Shown in figure 3 is a graphical representation of $s(x)$, 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.

To globally compare the two fuzzy entities *X* and *Y*, the maximum of *s* over *A* is taken: thus 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, which is called the *fuzzy equal*. In the previous picture, the value of *X* ~ *Y* is given by the ordinate of the marked point in figure 3.

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. However, speaking globally, the most likely situation corresponds to the greatest of these values. The equality of the two items may derive from their being both "male", or 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 credibility to the overall reliability. A prudent approach would, for the reliability of each case, establish the minimum of the two reliability coefficients, and the overall reliability would equal the greater 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 point marked in figure 3.

The definition of fuzzy equality is an example of generalization to fuzzy entities that are of familiar concepts, such as equality, counting, adding, averaging, and so on. We are not going to deal with these concepts any more: this paper shall only approach the counting of fuzzy quantities. To count occurrences, that is, to compute frequencies, we need to generalize the familiar operation of

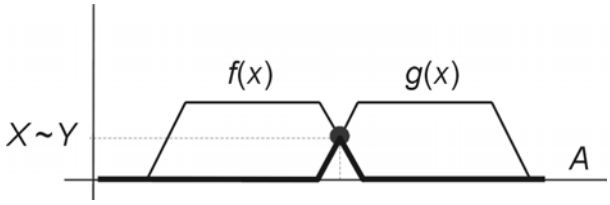

Figure 3. Graph of similarity function s *(heavy line) and the value of* X ~ Y *(marked point).*

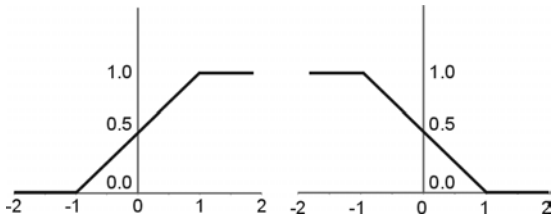

Figure 4: Derivation of m *(left, ordinates) and* f *(right, ordinates) from* k *(abscissa).*

counting: adding one when the desired result comes out (for instance, "female" when counting gender occurrences) and oppositely, adding zero 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 is in accord with common sense weighing in average evaluation and furthermore, also represents a particular case of the more general "Extension principle" (see Yager and Filev 1994:16-18).

5. Fuzzy entities in the case study

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

Gender may be considered 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. The other two attributes, however, require the definition of a function, as shown below.

There are several (in fact, infinitely many) possible ways to assign numerical values to the gender coefficients. The ones we chose are based on the following considerations:

- In this case study, few osteological coefficients (less than 20 %) go beyond +1 or –1; this can be considered the best possible result in these conditions.
- When the male coefficient gets the 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 the osteological coefficient by *k* and the male and female corresponding gender coefficients by *m* and *f* respectively, to derive *m* and *f* from *k* we can build the function shown in figure 4.

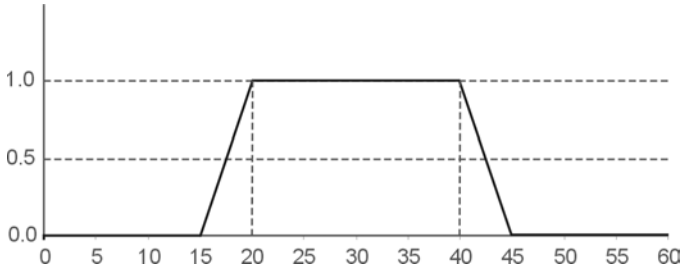

Figure 5: Fuzzy age coefficient (example for the age range 20 – 40).

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

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

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* remains undefined, since there were not enough elements 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 undecided regarding the known elements; hence, both gender are equally likely (or unlikely). The disadvantage lies in that the difference, if any, between the case of $k = 0$ and *k* not computable is lost; so one might prefer a different assignment as, for instance, $m = f = 0$. We believe that there is no relevant loss of information and furthermore, the method applies with both choices; so the one we adopted will have no consequence on the model's validity.

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

While the results of the second method are given in the form of a range with no other information included, the authors cited for the first one (Scarsini and Bigazzi 1995, 139-140), the central value *m* and the standard deviation *s*. So it is reasonable to assume that estimated ages have a Gaussian distribution, with mean *m* and standard deviation *s* as presented in the paper. Since the area below the Gaussian curve between $m – s$ and $m + s$ 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 $m – s$ and $m + s$ are the most probable and hence the most reliable, with a tail on both sides, having a lower probability. In terms of reliability of the result, we may therefore assume that this is the highest for the values within $[m – s, m + s]$, decreasing to zero outside; to keep things simple, the usual trapezoidal shape may be used, so that the reliability coefficient will be 1 within $[m – s, m + s]$, going to 0 at $m – s – q$ and at $m + s + q$, *q* being a positive number. In order to estimate *q*, the table of the normal distribution demonstrates that when $q = 0.5s$, 7% of the area is left out on each side; for $q = 0.64s$, the remainder is 5%; for $q = 1.33s$ 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 we cautiously chose $q = 0.5s$, considering those cases that have a probability less than 0.07 as unreliable. Since the age range has a width of $2s$, 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.

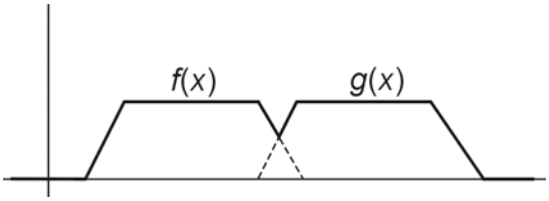

Figure 7: Graph of fuzzy OR (heavy line).

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.

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

When the osteological investigation renders two distinct, not overlapping ranges for the age, they will correspond to a fuzzy reliability function that is built taking into consideration the two separate parts as distinct reliability functions $f(x)$ and $g(x)$, and defining their fuzzy OR as follows:

$$f \text{ OR } g(x) = \max\left(f(x), g(x)\right)$$

with the graph shown in figure 7.

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

Let us consider two cases. The first case incorporates an osteological age range of 22 – 30 years and the second 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: our cautious assumption determines 2 years for the first one and 4 years for the second one (traditional statistical assumptions would have fixed them at 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. Now who is "younger"?

This simple example shows that our common sense reasoning may be fallacious and new categories need to be introduced, even for

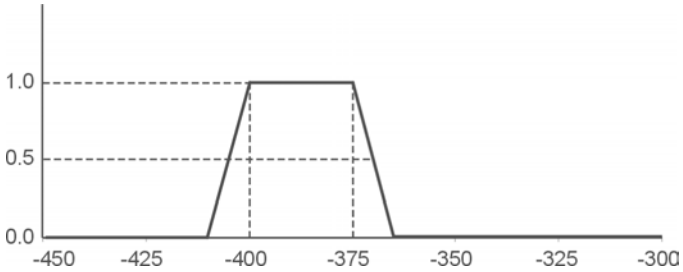

Figure 8: Fuzzy chronology with a slack of 15 years (example).

simple comparisons, which lose any significance when applied to statistical data.

As far as chronology is concerned, nominal constants were used in this context, as usual; for instance, the "first quarter of the 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 *d* on each side, on which the fuzzy reliability coefficient varies uniformly from 0 to 1, or vice versa.

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

It might be argued that wider intervals lower the confidence level in each single value of a fuzzy value. For instance, an age interval such as 20 – 40 would make any single age, 35 for instance, less credible. If this is the case, and confirmed by the osteological method used to determine the age interval, the top value of the reliability function should be lowered accordingly. However, provided with no other information, then all values should have equal and top reliability factors, 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 their number since there is no constraint when adding up to 1. This substantiates our use of this approach as opposed to 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, rather with "exact" figures having a different reliability, so as to establish archaeological inference knowing how reliable they are.

Concluding this paragraph on the evaluation of fuzzy reliability coefficients, we want to emphasize 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 proved beyond doubt that any attempt to construct a completely deterministic model, intended to provide archaeological results only by means of computations, would be 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 data in a database, it is necessary to create special data types corresponding to fuzzy entities, which are fuzzy labels and fuzzy values, and to set the rules to process them. Since these have already been defined in Crescioli et al. (2000), 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, possibly with different coefficients. If we agree to always put the nominal elements in the same (not relevant) order in the array, then the fuzzy label is characterized by the domain: that is, the common set of *n* possible labels and an *n* multiple of numbers, the values of which are different for each instance of the label. In our case, the fuzzy gender is characterized by the domain as follows: the two labels "male" and "female", and a couple of numbers as (0.3, 0.7), having conventionally agreed that the former refers to "male" and the latter to "female".

The definition of each member of the data type FUZZY_LABEL, and likewise also FUZZY_GENDER, thus requires that it store the domain somewhere, being {"male", "female"} in this case. It then consists of a one-dimensional array of real numbers.

Regarding fuzzy values, the required 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 et al. (2000). Our models set 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 portraying 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 in order to make the use of constants in queries easier, they are denoted as functions with no parameters, for instance YOUNG().

To determine the values of these fuzzy constants, we refer to the commonly used ranges by anthropologists; so that, for instance, YOUNG() means an age included within the range of 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. An example: the comparison be-

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 1: Results of FUZZY_AGE ~ YOUNG() (example).

tween the FUZZY_AGE attribute 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 1, where the descriptions are presented instead of the values of fuzzy entities in order to ease readability.

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 PostgreSQL for this, a RDBMS available under a Linux operating system. Furthermore, being 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, an SQL environment command line 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 for data input, as direct conversion was quicker. The data were then manipulated to give expressions such 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 in 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 and not overlapping intervals of 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 the table TOMB necessitated 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.

The GRAVEGOODS table was created in a similar manner with data concerning grave goods.

The constants were inserted into the CONSTANTS table and then they used to create functions and operators, such as fuzzy equality ~ . 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 was introduced in order to allow an expression in 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}').

9. Archaeological application to the case study

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

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

Within the framework of demographic applications to cemetery interpretation, two criteria are particularly significant for the verification of how representative the funerary sample actually is (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 preindustrial 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, then 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 emphasized more or less according to the economic and social struc-

Land property	Maisto-Boccia	Rossomando	Tascone-Di Dato	Cemetery
Age ratio (Serritella's data)				
Infant	22.9%	33.3%	16.0%	22.7%
Children – Young	0.0%	6.7%	0.0%	1.3%
Adult – Elderly	54.3%	53.3%	76.0%	61.3%
Indeterminate	22.8%	6.7%	8.0%	14.7%
Total	100.0%	100.0%	100.0%	100.0%
Ratio A/(I+C)	2.37	1.33	2.75	2.56
Age ratio (Fuzzy model)				
Children	43.9%	45.3%	28.5%	41.0%
Adult	56.1%	54.7%	71.5%	59.0%
Ratio A/C	1.28	1.21	2.50	1.44
Gender ratio (Serritella's data)				
Male	47.4%	75.0%	63.1%	58.7%
Female	42.1%	25.0%	31.6%	34.8%
Indeterminate	10.5%	0.0%	5.3%	6.5%
Ratio M/F	1.13	3.00	2.00	1.69
Gender ratio (Fuzzy model)				
Male	53.1%	70.4%	57.6%	57.2%
Female	46.9%	29.6%	42.4%	42.8%
Ratio M/F	1.13	2.38	1.36	1.34

Table 2: Age and gender ratio corresponding to land property (percentages). Derived from Serritella (1995: 116, 121, 123) and computed from the fuzzy model.

ture of the community, becoming inadequate when egalitarian ideologies prevail.

In the past, applications of statistical methods to the study of funerary custom have been based on purely quantitative logic. Mathematical models were used to measure - "objectively" - the richness of a tomb, to determine the funerary variability or to identify social hierarchy on the basis of "energy expenditure" (see Cuozzo 1994:268, Cuozzo 1996). Contrarily, we suggest the use of fuzzy logic (using palaeo-anthropological demographic data) to estimate how much a funerary sample is representative of the community. A few examples are given below, and we consider this perspective a potentially substantial contribution for the execution of an analysis that aims to determine 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 true ratios of children to adults or males to females, thus turning awry any deduction derived from the sample.

Consider the first block of table 2, derived from those published in Serritella (1995: 116, 121, 123), counting the occurrences of age categories and shown as percentages of the total (blocks are identified by heavy lines). The tombs are grouped by modern landowners, 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. Serritella demonstrated that this grouping reflects significant differences in chronology or rite.

Indeterminate gender assignments add up to 15 % of the total and consequently, they may significantly alter the confidence level 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 highly representative one.

Using fuzzy coefficients, the fuzzy age "Infant or Child" class, ranging from 0 to 20 years, needs to be introduced, as well as the "Adult or Elderly" class, ranging from 21 upwards. The data then needs to be compared with the two new constants. We then compare Serritella's results with those shown in the second block of table 2, which are easily obtainable by means of an SQL query on the database. The "Rossomando" and "Maisto-Boccia" groups fit the model, while the "Tascone-Di Dato" does not. The latter shows a prevalence of adult burials. Considering the male to female ratio, the values shown in the third block of table 2 are from Serritella's work and the ones shown in the last block of the same table are from our database. From this it follows that the "Maisto-Boccia" group portrays a representative sample, the "Tascone-Di Dato" is less so and the "Rossomando" group is not.

After 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. An explanation for this discrimination, in terms of social status and age and gender roles, can derive only from an investigation of the grave goods; that is, to understand the underlying cultural model of social representation. However this reaches beyond the aim of the present paper. Comparing our results with Serritella's, what she suggested is confirmed by our own research; however, with a much higher level of reliability, due to the absence of indeterminate cases, which in her tables should suspend every conclusion.

10. Conclusions

If the proposed model is accepted, at least in a negative sense, if not positive, then greater caution should be instigated when using statistical data in archaeological investigations. It will no longer be possible always to take the reliability of archaeometric data for granted; at least, not when these data are critical for the validity of an interpretation model. Nevertheless, we hope that our model, together with the computer tools we made available (possibly improved by future work) will help research. The user-friendly aspect of computers and the extensive availability of software tools

together have the positive effect of spreading the advantages of their applications, while at the same time their usage occurs unawares. Perhaps this contribution will indeed augment the awareness of archaeologists 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 Barceló, J.A., Forte, M., Sanders, D.H. (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., 2000. 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 necropolis. In Francovich, R. and Manacorda, D. (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 Barcelò, J.A., Briz I. and Vila, A. (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 theory: 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 theory and applications. In Lock, G. and Stančič, Z. (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 Dingwall, L., Exon, S., Gaffney, V., Laflin, S., and Van Leusen, M. (eds.), *Computer Applications and Quantitative Methods in Archaeology 1997*: 19-27. Oxford: Archaeopress (BAR International Series 750).