Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2008/01

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Identification of the Victims of a Mass Fatality Incident based on nuclear DNA evidence

David Cavallini, Fabio Corradi

Università degli Studi
di Firenze

Identification of the Victims of a Mass Fatality Incident based on nuclear DNA evidence

By David Cavallini and Fabio Corradi

Department of Statistics, University of Florence

50134 Florence, Italy

cavallin,corradi@ds.unifi.it

Summary

This paper focuses on the use of nuclear DNA Short Tandem Repeat traits for the identification of the victims of a Mass Fatality Incident. The goal of the analysis is the assessment of the identification probabilities concerning the recovered victims. Identification hypotheses are evaluated conditionally to the DNA evidence observed both on the recovered victims and on the relatives of the missing persons disappeared in the tragical event. After specifying a set of conditional independence assertions suitable for the problem, an inference strategy is provided, treating some points to achieve computational efficiency. Alternative solutions to the problem will also be illustrated for comparison purposes. Finally, the proposal is tested through the simulation of a Mass Fatality Incident and the results are compared with the considered alternative solutions.

1. Introduction

Terroristic attacks, natural calamities and transportation crashes have in the recent years caused a relevant number of Mass Fatality Incidents (MFI), posing challenging identification problems to the authorities. In such circumstances, identification has traditionally been attempted through the bodies' direct recognition and by a comparison of some victims' characteristics to the relevant records of missing persons presumably involved in the tragical event.

Often, little but some biological material can be recovered from the victims and several DNA Short Tandem Repeat (STR) loci can be employed to attempt identification. In such cases, the identification process does not necessarily require the missing persons' biological samples, since, exploiting DNA heritability, some genetic material obtained from their relatives can be used instead.

To find a specific missing person among the victims, Clayton et al. (1995) and Cash et al. (2003) evaluated as many likelihood ratios (LR) as the number of recovered bodies. Each LR was separately assessed as the probability to observe a victim and the missing-related evidence, conditionally to a pair of competitive hypotheses. The first conjecture reckons that the victim is the missing individual; the alternative assumes that the missing person is not related to the victim, being this latter a generic member of a specified genetic population. The method is strictly derived from the widely accepted solution of common identification problems in paternity cases. There, an individual is alleged to be in a certain position in a pedigree and no one else is alternatively specified. The approach, named kinship analysis by Brenner (1997), if repetitively applied in a MFI setting, does not provide encouraging results: in fact, often, for each missing person, some large LRs are obtained with respect to different victims, not leading to conclusive results in terms of identification. False positives were justified in Brenner & Weir (2003) by the consideration that the expected number of individuals whose genetic profile is compatible with the unobserved missing person's one increases according to the population size and this is not negligible. In the same work the authors also recognized that: "the need to consider not only the evidence from the similarity of a victim sample to a particular family, but also its dissimilarities to other families, is overlooked".

This is exactly our opinion: the poor result obtained using STR DNA evidence was due to an improper definition of the alternative hypothesis, which is not constituted by the generic member of the genetic population but must contemplate all the recovered and the unrecovered victims.

A step ahead has recently been suggested by Brenner (2006), who proposed to consider at the same time all the missing individuals occurring in each familial group. However, if this approach opens the way to identification when no genetic data are available from the relatives of the missing persons, the families are still considered separately.

Recently, Cavallini & Corradi (2006), considering another identification problem, introduced the possibility to evaluate the probability that a trace of unknown origin, a victim here, could be a certain individual, compared to the alternative that considers other candidates to the identification, both observed and unobserved. This approach will be considered later and referred as the *Victim model*.

The identification of a victim starting from the hypothesis that he/she is one of the missing persons is only one of the two possible ways to pose the problem, the other is to start from the hypothesis that the missing person is one of the recovered victims. This point of view will also be considered and named the *Family model*.

The proposal we describe consists in considering simultaneously the probabilistic identification of all the victims and the probabilistic assignment of the recovered bodies to the missing persons. For this reason we have named the proposal the *Complete model.*

Finally, even if we are aware of the many other important questions implied in a MFI identification process, like duplication of the traces or the assessment of the true familial relationship, this paper concentrates only on the probabilistic issues strictly related to the treatment of DNA evidence, referring for detailed practical considerations to Cash et al. (2003).

2. Basic Ingredients

Let N the number of persons involved in a MFI. Temporarily assume N exactly known, as it happens when an aircraft accident occurs and the passenger and the crew lists are available. If less information is available, N can be reckoned at a large conservative value or assessed probabilistically.

The aim is to identify the members of the set of recovered victims set $\mathcal{V}$ by means of the missing individuals in the set $\mathcal{M}$. The latter are assumed to belong to families placed in the set $\mathcal{F}$. Missing individuals in each family are posed in the set $\mathcal{M}_f$, besides the observed individuals $\mathcal{O}_f, f \in \mathcal{F}$. Referring to a certain family, the specification of a pedigree often requires the specification of some unobserved family' members who are posed in the set $\mathcal{U}_f$. Also $\mathcal{M} = \bigcup_{f \in \mathcal{F}} \mathcal{M}_f$, $\mathcal{O} = \bigcup_{f \in \mathcal{F}} \mathcal{O}_f$, $\mathcal{U} = \bigcup_{f \in \mathcal{F}} \mathcal{U}_f$.

From now on, $n(\cdot)$ indicates the cardinality of the set in the argument; $I_A(B)$ is the usual indicator function, which is 1 if $A = B$ or it is 0 otherwise; P_n, $D_{n,k}$ and $C_{n,k}$ indicate, respectively, permutations, dispositions and combinations of k elements taken from a set of cardinality n.

We define the identification hypothesis, H, taking into account all possible ways in which the missing individuals can identify the recovered victims. Once the support provided by all the evidence to each state of the identification hypothesis has been calculated, the probabilities evaluating the assignment of the victims to each of the missing persons or, conversely, the identification alternatives for each recovered victim, can be obtained from a simple marginalization procedure.

To formalize the possibility that not all the victims have been recovered, start augmenting $\mathcal{V}$ by a ?, a generic unrecovered victim, so that $\mathcal{V}^* = \mathcal{V} \cup \{?\}$. Define $H_m = v \in \mathcal{V}^*$, $m \in \mathcal{M}$, the hypothesis random variable concerning the assignment of one of the members of the set $\mathcal{V}^*$ to the m-th missing person. If the m-th missing person is considered in isolation, the corresponding identification random variable can assume values in $\mathcal{V}^*$ without constraints.

Instead, if more than one H_m are considered jointly, a multivariate random variable H must be defined such that multiple assignments of the same victim to different missing persons are not allowed. Let H^t, a generic configuration, conform to the mentioned constraint and formally defined by:

$$H^t = \{H^t_m : m \in \mathcal{M}\} \text{ where, } H^t_s =? \text{ or } \forall g \neq s \; H^t_g \neq H^t_s \; . \qquad (1)$$

If the number of the recovered victims is equal to the number of the individuals involved in the disaster, $n(\mathcal{H}) = P_N$ since each victim can be identified by only one missing person; otherwise, if $n(\mathcal{V}) < N$, then $n(\mathcal{H}) = D_{N,n(\mathcal{V})}$.

The individuals implied in the analysis are considered only with respect to nuclear STR DNA *loci*, those commonly used for forensic identification. We do not refer to a particular set of them since our findings are independent of such choice.

In a locus we observe a genotype, i.e. two alleles inherited from the father and the mother even if their origin is not recoverable. A random variable X represents the uncertainty about genotypes and a determination of X is simply indicated by x. The X probability function can be provided by two kinds of models:

- Segregation models: for a locus, they evaluate the probability of an offspring's genotypes conditionally to their parents. The first mendelian law specifies the genotype' s probability of a child, c, given the genotypes of their parents, m and f. If $x_c = (t, z)$, $x_m = (i, j)$ and $x_f = (r, s)$, we have:

  $$Pr(x_c \mid x_m, x_f) = \frac{1}{4}(I_{\{i,r\}}(x_c) + I_{\{i,s\}}(x_c) + I_{\{j,r\}}(x_c) + I_{\{j,s\}}(x_c)). \quad (2)$$

  If mutations or laboratory errors are involved, other more sophisticated models are required to describe the segregation process Dawid et al. (2007).

- Population models: they determine the probability of an individual's genotype conditionally to his \ her belonging to a specified population in which the alleles' probabilities, θ, are assumed known. The most popular of such models derives by the conditions introduced by Hardy-Weinberg for a population in equilibrium, Weir (1996). In this case the genotypic probability is calculated from the probabilities of the alleles in the population. For a generic individual m, the genotype probability is:

$$Pr(x_m = (i,j) \mid \theta) = \theta_i \cdot \theta_j \cdot (1 + I_{\{i,j:i \neq j\}}\{i,j\}), \tag{3}$$

If required, also in this case it is possible to make use of more sophisticated models, as described in Evett & Weir (1998), to take into account possible inbreeding and coancestry characteristics in the populations.

As a matter of notation $X^V = \{X_v^V : v \in \mathcal{V}\}$ refers to the recovered victims genotypes; $X^F = \{X_f^F : f \in \mathcal{F}\}$ regards the families to which the missing persons belong and can be split into $X_f^F = \{X_f^M, X_f^O, X_f^U\}$, according to the family's members introduced in section (3). We also define, with X^M, the set of all the missing persons' genotype random variables and, with X^O and X^U, the sets of their observed and unobserved relatives.

3. The complete Model

To make inference about H consider the following decomposition of the joint probability distribution of the random variables implied in the analysis:

$$Pr(X^V, X^F, H) = Pr(X^V \mid X^F, H)Pr(X^F \mid H)Pr(H). \tag{4}$$

Each factor in (4) can be simplified by some conditional independence assertions.

a) $X^F \perp\!\!\!\perp H$, i.e. the identification hypothesis does not modify the probabilistic relations among the genotype random variables of the familial groups asking for the identification of their members. This implies:

$$Pr(X^F \mid H) = Pr(X^F) \tag{5}$$

b) Familial groups are defined to that they include all the observed and unobserved individuals known to be related. Two families cannot share their members, otherwise they are merged. This implies that the random variables related to the genotypes of individuals belonging ti different families are independent:

$$Pr(X^F) = \prod_{f \in \mathcal{F}} Pr(X^F_f). \tag{6}$$

c) To decompose $Pr(X^V \mid X^F, H)$ consider that, $\forall t$, $\exists!\ m$ such that $H^t_m = q \in \mathcal{V}$. This implies $X^V_q \equiv X^M_m$, providing $X^V \perp\!\!\!\perp X^O, X^U \mid X^M, H$, so that:

$$Pr(X^V \mid X^F, H) = Pr(X^V \mid X^M, H). \tag{7}$$

A formal expression of the likelihood of the observed evidence, $X^V = x^V, X^O = x^O$, conditionally to each of the H^t states, derived from (4), (5) and (7), is:

$$Pr(x^V, x^O \mid H^t) = \sum_{X^M, X^U} Pr(x^V \mid X^M, H^t) Pr(x^O, X^U, X^M). \tag{8}$$

To evaluate (8), define as $\mathcal{M}_f^t = \{m \in \mathcal{M}_f : H_m^t \in \mathcal{V}\}$ the sets of missing persons in the families having victims assigned by a specified H^t, being these families posed in the set $\mathcal{F}^t = \{f \in \mathcal{F} : \mathcal{M}_f^t \neq \emptyset\}$. Also let $X_f^{M_t} = \{X_m : m \in \mathcal{M}_f^t\}$ the random variables of the missing persons' genotype in the f-th family, being $X_f^{V_t}$ the matching victims' genotypes assigned by H^t, so that:

$$Pr(x^V \mid X^M, H^t) = \prod_{f \in \mathcal{F}^t} Pr(x_f^{V_t} | X_f^{M_t}, H^t), \tag{9}$$

and:

$$Pr(x_f^{V_t} | X_f^{M_t}, H^t) = \begin{cases} 1 & \text{if } X_f^{M_t} = x_f^{V_t} \\ \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Taking account of (6) and (9), $\forall t$, the likelihood can be factorized as follows:

$$\begin{aligned} Pr(x^V, x^O | H^t) = & \prod_{f \in \mathcal{F}^t} \sum_{X_f^M, X_f^U} Pr(x_f^{V_t} | X_f^{M^t}, H^t) Pr(x_f^O, X_f^U, X_f^M) \\ & \cdot \prod_{f \in \mathcal{F} \setminus \mathcal{F}^t} \sum_{X_f^M, X_f^U} Pr(x_f^O, X_f^U, X_f^M). \end{aligned} \tag{11}$$

Then, by (10):

$$\sum_{X_f^M, X_f^U} Pr(x_f^{V_t} | X_f^{M_t}, H^t) Pr(x_f^O, X_f^U, X_f^{M_t}) = Pr(x_f^O, X_f^{M_t} = x_f^{V_t}), \tag{12}$$

which is equivalent to a transfer of evidence from the victims to the corresponding missing individuals. Finally, the likelihood results to be:

$$\begin{aligned} Pr(x^V, x^O | H^t) &= \prod_{f \in \mathcal{F}^t} Pr(x_f^O, X_f^{M_t} = x_f^{V_t}) \prod_{f \in \mathcal{F} \setminus \mathcal{F}^t} Pr(x_f^O) \\ &\propto \prod_{f \in \mathcal{F}^t} \frac{Pr(x_f^O, X_f^{M_t} = x_f^{V_t})}{Pr(x_f^O)}, \end{aligned} \tag{13}$$

where the final expression is obtained by dividing for $\prod_{f \in \mathcal{F}} Pr(x_f^O)$, a quantity independent of H^t.

A more intriguing formulation of the likelihood for H^t is possible if all the missing persons belong to the same genetic population. In such case, (13) can be divided by the probability to observe the recovered victims, assuming that they belong to the considered genetic population, so that:

$$\begin{aligned} Pr(x^V, x^O | H^t) &\propto \prod_{f \in \mathcal{F}^t} \frac{Pr(X_f^{M_t} = x_f^{V_t}, x_f^O)}{Pr(x_f^O) \prod_{m \in \mathcal{M}_f^t : H_m^t = v} Pr(X_m^{M_t} = x_v^{V_t})} \\ &\propto \prod_{f \in \mathcal{F}^t} \mathbb{LR}^t(f). \end{aligned} \tag{14}$$

This representation shows how the likelihood for H^t can be expressed by the likelihood ratios $\mathbb{LR}^t(f), f \in \mathcal{F}^t$, the usual result of a kinship analysis. More specifically, each $\mathbb{LR}^t(f)$ is the ratio between the probability of the familial observed evidence if the missing individuals are the assigned victims and the probability of the evidence obtained evaluated according to the hypothesis that the recovered victims are not the families' missing individuals but simply belong to the relevant genetic population.

The expression of the likelihood in terms of $\mathbb{LR}$ has the advantage to allow for separate computations at familial level, paving the way to parallel calculus strategies. Furthermore, (14) points out which families potentially have a $\mathbb{LR} \neq 1$, excluding those which structurally cannot provide information to the hypothesis, always showing a $\mathbb{LR} = 1$.

A noticeable case arises if, in a familial group, the relationships are known but no familial evidence is available *and* more than one missing individual perished in the MFI. If a certain H^t assigns more than one victim to the family, $\mathbb{LR} \neq 1$, since the probability to observe the victims - evaluated assuming the familial relationship - differs if the assumption of independence holds.

Among non informative families, unclaimed missing persons are a typical example. The case is formally represented by a missing individual searched by an empty family so that, if a victim is assigned, the corresponding $\mathbb{LR}$ is equal to one. This consideration has two important consequences.

First of all we can restrict the computation of (14) only to the potentially informative families and the associated missing individuals, respectively defined by:

$$\begin{aligned} \mathcal{F}^* &= \{f \in \mathcal{F} : n(\mathcal{M}_f) > 1 \text{ or } \mathcal{O}_f \neq \emptyset\} \\ \mathcal{M}^* &= \{m \in \mathcal{M}_f : \mathcal{F} \in \mathcal{F}^*\}, \end{aligned} \tag{15}$$

being the complementary set of the non informative families defined by:

$$\begin{aligned} \mathcal{F}^+ &= \{f \in \mathcal{F} : n(\mathcal{M}_f) = 1 \text{ and } \mathcal{O}_f = \emptyset\} \\ \mathcal{M}^+ &= \{m \in \mathcal{M}_f : \mathcal{F} \in \mathcal{F}^+\}. \end{aligned} \tag{16}$$

It follows that, for a given H^t, not all the families contribute to the likelihood (14) but only those in the set $\mathcal{F}^t \cap \mathcal{F}^*$, so that the likelihood can be written as:

$$Pr(x^V, x^O | H^t) \propto \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} \mathbb{LR}^t(f). \tag{17}$$

The second important consequence is that many configurations differ only for the victims allocated in $\mathcal{F}^+$, so they have the same likelihood.

Formally, if $H^t \neq H^s$ but $\mathcal{F}^t \cap \mathcal{F}^* = \mathcal{F}^s \cap \mathcal{F}^*$, then

$$Pr(x^V, x^O|H^t) = Pr(x^V, x^O|H^s). \tag{18}$$

Since the goal of the analysis is to provide inference on the identification hypotheses concerning the members of the set $\mathcal{M}^*$, it is convenient to partition each H^t accordingly. So we have $H^t = [H^t_*, H^t_+]$, being $H^t_* = \{H^t_m : m \in \mathcal{M}^*\}$ of real interest and $H^t_+ = \{H^t_m : m \in \mathcal{M}^+\}$ a nuisance random vector.

If, for $t \neq s$, (18) holds, these configurations belong to the same inferential class. It is computationally convenient to evaluate the classes' cardinality since inferring on the hypotheses concerning the $\mathcal{M}^*$ members, the contribution of each class is simply equal to its cardinality times its members' likelihood.

To evaluate the classes' cardinality consider that, if two configurations are in the same class, they have $H^t_* = H^s_*$ and $H^t_+ \neq H^s_+$. So, how many members are in the class depends on the number of ways H_+ can appear, i.e. on the possible assortments of the victims allocated among the $\mathcal{M}^+$ members. If i^t is the number of victims assigned by H^t_*, then the class at which the t-th configuration belongs has cardinality $D_{n(\mathcal{M}^+),n(\mathcal{V})-i^t}$.

To produce inference on hypotheses concerning the members of $\mathcal{M}^*$, it is convenient to define a new hypothesis random variable, H_*, concerning exclusively the members of $\mathcal{M}^*$. Let H_*^t a generic configuration characterizing an inferential equivalent class, formally defined by:

$$H_*^t = \{H_m^t : m \in \mathcal{M}^*\} \text{ where, } H_s^t =? \text{ or } \forall g \neq s \; H_g^t \neq H_s^t. \tag{19}$$

If a uniform prior was posed on the H^t's, i.e. no information is assumed on the identity of the recovered victims, inference on H_*^t can be obtained by marginalizing with respect to H_+^t, thus obtaining:

$$Pr(H_*^t|x^O, x^V) \propto D_{n(\mathcal{M}^+),n(\mathcal{V})-i^t} \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} \mathbb{LR}^t(f). \tag{20}$$

To appreciate the saving in computational efforts, note that the cardinality of H_* can be evaluated and compared with $n(\mathcal{H})$.

The cardinality of $\mathcal{H}^*$ can be evaluated defining $i \in \mathcal{I}$ as the number of possible victims allocated to the $\mathcal{M}^*$, with $\mathcal{I} = \{max(0, n(\mathcal{V}) - n(\mathcal{M}^+)), \ldots, min(n(\mathcal{M}^*), n(\mathcal{V}))\}$. For each i the number of possible equivalent classes is $C_{n(\mathcal{V}),i} \cdot C_{n(\mathcal{M}^*),i} \cdot P_i$, so that:

$$n(\mathcal{H}^*) = \sum_{i \in \mathcal{I}} C_{n(\mathcal{V}),i} \cdot C_{n(\mathcal{M}^*),i} \cdot P_i. \tag{21}$$

When the missing individuals belong to more than one population, inference requires more efforts. Actually, the probability for the victims to simply belong to the specified genetic population, introduced to achieve (14), now varies from a configuration to another, depending on which genetic population belong the missing persons who have victims assigned.

To take account of the population variety, introduce the set $\mathcal{K} = \{1, \ldots, k\}$, containing the population labels and let $\Pi = \{\pi_i : i = 1, \ldots, k\}$ be the proportions of missing individuals belonging to each population. Also, let $G_m = i \in \mathcal{K}$ the indicator random variable assigning the m-th missing person to the i-th genetic population, being $G = \{G_m : m \in \mathcal{M}\}$.

Now we re-derive the likelihood from the first line of (13), splitting the product into informative and non informative families:

$$Pr(x^V, x^O | H^t) \propto \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} \frac{Pr(x_f^O, X_f^{M_t} = x_f^{V_t})}{Pr(x_f^O)} \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^+} Pr(X_f^{M_t} = x_f^{V_t}). \quad (22)$$

If we multiply and divide (22) by the probability to observe the victims, arranged according to $\mathcal{F}^*$ and $\mathcal{F}^+$, we get the likelihood expression:

$$
\begin{aligned}
Pr(x^V, x^O | H^t) \propto & \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} \mathbb{LR}^t(f) \prod_{m \in \mathcal{M}_f^* : H_m^t = v} Pr(X_m^{M_t} = x_v^{V_t}) \\
& \cdot \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^+} \prod_{m \in \mathcal{M}_f^+ : H_m^t = v} Pr(X_m^{M_t} = x_v^{V_t}),
\end{aligned}
\tag{23}
$$

where the likelihood ratios for the informative families in (20) still appear but the probability to observe the victims now depends on H^t, and becomes informative too.

Now consider the marginalization procedure required to obtain inference about H_*. Similarly to the previous case, the probability to observe the victims assigned to the members of $\mathcal{M}_f^*$ does not vary; on the opposite, depending on the elements of $\mathcal{M}^+$ to which the $n^t = n(\mathcal{V}) - i^t$ victims are assigned, this probability varies according to the population which the unclaimed missing individuals belong to. This fact obviously ruins the idea of inferential equivalent classes but it is still convenient to express the likelihood, related to each different H_*^t, by means of a single expression. This is obtainable considering all the possible ways the n^t unclaimed victims can be allocated among the populations and the joint assignment probability G, finally providing the required marginalization.

To achieve this result, first consider the number of unclaimed missing individuals in each population,

$$N_i^+ = N\pi_i - \sum_{f\in\mathcal{M}_f^*}\sum_{m\in f} I_{\{i\}}(G_m),\ \forall i \in \mathcal{K}, \tag{24}$$

and their total number,

$$N^+ = \sum_{i\in\mathcal{K}} N_i^+, \tag{25}$$

two quantities not depending on the configurations.

Once an H_*^t has assigned i^t victims among the $\mathcal{M}^*$ missing individuals, the remaining n^t have potentially $(n^t)^k$ ways to belong to the k different populations even if not all the population assignments are allowed, since $\forall i, \sum_{f\in\mathcal{M}_f^+}\sum_{m\in f} I_{\{i\}}(G_m) \leq N_i^+$.

For every arbitrary order of the n^t victims, the joint probability of the G indicator random variables depends on the $N_i^+, i = 1, \ldots, k$ and on N^+; moreover if G is decomposed accordingly to the telescopic rule, and g_{-m} indicates the population assigned to the first $m-1$ missing persons, it can be shown that, for every H^t belonging to a specific equivalent class:

$$Pr(G) = \prod_{m\in\mathcal{M}_f^+ : H_m^t = v} Pr(G_m | G_{-m} = g_{-m}) = \frac{\prod_{i=1}^k D_{N_i^+, n_i^t}}{D_{N^+, n_t}}, \tag{26}$$

where, according to the order of the set $\mathcal{M}^+$, g_{-m} indicates the values assumed by $\wedge_{m=m+1}^{n^t} G_m$ random variables, being n_i^t the victims assigned by the H^t to the i-th population. If, again, on the H^t a uniform prior is posed, inference on H_*^t can finally be derived from:

$$
\begin{aligned}
Pr(H_*^t|x^V, x^O) \propto \prod_{f\in\mathcal{F}^t\cap\mathcal{F}^*} \mathbb{LR}^t(f) \prod_{m\in\mathcal{M}_f^*:H_m^t=v} Pr(X_m^{M_t}=x_v^{V_t}) \\
\cdot \sum_{G_1}\cdots\sum_{G_{n^t}} \left[\left(\prod_{m\in\mathcal{M}_f^+:H_m^t=v} Pr(X_m^{M_t}=x_v^{V}|G_m)\right) Pr(G)\right],
\end{aligned}
\tag{27}
$$

which represents the generalization of (20) to k populations.

4. An alternative inferential strategy

Starting from the seminal paper of Dawid et al. (2002), an increasing number of complex identification issues based on DNA evidence has recently been dealt with by means of Probabilistic Expert Systems (PES), see for example Mortera et al. (2003), Cavallini & Corradi (2006) and Vicard et al. (2007).

A PES, Cowell et al. (1999), is constituted by a Directed Acyclic Graph (DAG) and a set of conditional probability distributions, one for each of the DAG nodes. Nodes represent the relevant random variables for the problem at hand and some of them are suitably linked by arrows or directed edges. The term PES emphasizes that the required conditional probability distributions are provided by experts, who exploit results achieved in many sciences with a view to provide a coherent system of knowledge. The main aim of a PES is to get, efficiently, the posteriors of some interesting and unobservable random variables; this goal differs from the one of a Bayesian Network, which is mainly interested in learning the structure and parameters of the stochastic system.

The main advantage of a PES is the possibility to include a detailed description of many sources of uncertainty obtaining inference numerically, i.e. without deriving inference analytically. In a PES, exploiting conditional independence relations, inference is performed for groups of random variables; the results are then conveyed among groups through *separators*, i.e. nodes belonging to more than one group. A comparison between the analytical and graphical approaches to inference about identification can be appreciated likening how the issue is handled for DNA mixed profiles by (Evett & Weir (1998), cap. 7) and (Mortera et al. (2003)).

Now we turn our attention to the representation of the proposed *Complete Model*, Fig. (1). The graph embeds the conditional independence assertions implied in (6) and (7) and all the relations among missing individuals, their relatives, the victims and the identification hypothesis.

Each graph represents a family, not being displayed the specific relationships among the missing individuals and their relatives. An example of a possible box content is in Fig. (4). Although this representation can produce, by means of a propagation algorithm, the same numerical results as (21) or (27), computations are possible only if a small number of victims and/or missing persons is involved. The difficulty is due to the ambition of the model, which requires to evaluate *and* compare simultaneously all the possible ways to assign the recovered victims to the missing persons. This implies an extremely large number of states for the hypothesis random variable; also, the conditional probability tables (CPT) for each X^V rapidly increase in dimension, depending on the number of H states and of missing persons. Making computations with CPTs of such dimension, is a task beyond customary computational resources and this motivates our efforts to derive inference from a familial base. Expressing the *Complete model* by a PES, not only produces a vivid representation of the relationships among the random variables according to the assertions of conditional independence but it also allows to easily derive the *Victim model* and the *Family model* introduced in section (1).

The *Victim model*, Fig. (2), restricts the victim set to only one victim at a time, attempting identification through searching among the missing persons. This model is essentially the one proposed, for another identification issue, by Cavallini & Corradi (2006), who also provided a way to manipulate the network so that standard propagation algorithms were able to derive inference.

The *Family model*, Fig. (3), only considers one request from family at a time, trying to find, among the victims, the individuals lost by the family. This approach is similar but not identical to the model proposed by Brenner (2006) since he introduced, separately, all the possible subsets of victims, as candidates to identification and not jointly as we propose.

5. A simulation-based example

In this section we desplay the results of an identification process carried on 14 individuals belonging to 10 families and disappeared in a simulated MFI. The example has been expressly built to concentrate, on a manageable number of individuals and families, many of the difficulties which typically arises in the field. The choice to simulate data gives us the opportunity to check the results, since we known the identity of the victims.

In each family, individual genotypes posed on the roots of the graph are simulated from a genetic population; the remaining nodes are derived from the first mendelian law. The missing persons' DNA profiles are removed from the families they belong to and posed in the victims' data set, canceling their identity.

In the example, three families, labeled 1, 7 and 10, look for only one missing person posed in direct lineage with the claiming relatives; in other four families, 3, 4, 5, and 6, one sibling is looking for another sibling; the remaining three families, 2, 8 and 9, search more than one of their members, but only the relationship among the missing individuals is assumed known, not being available any familial donor.

Starting from uninformative prior probabilities on every possible configuration H, the exercise consists in assessing the identification probability making use of the *Victim*, the *Family* and the *Complete* models. For each missing person we provide the posterior probabilities to find the corresponding victim or, if this latter has not been recovered yet, the probability that none of the recovered victims is the missing person.

Three different stages of the identification process are conceived. The first one figures out the availability of all the victims' profiles and of some pieces of information about all the missing persons. This setting is called the *Final stage* of the identification process. Reducing the number of the recovered victims and of the claimed missing persons, we provide the *Intermediate stage*. Finally a further decrease in information produces the *Early stage*.

The synthesis of the *Final stage* state of information is in Table (1), the labels explained in Fig. (4) and the models' results in Table (2).

First, note that every time the observed relatives and the claimed missing persons are in a direct lineage, all the models provide a high posterior probability of correct identification. This happens because the first Mendel law produces a large number of exclusions, eliminating all the victims incompatible with the relatives in the direct lineage. On the contrary, for those cases where one sibling is looking for another sibling, the *Victim Model* is not always successful. For instance, although the siblings in the 5-th family share one allele on many loci (data not displayed), since these alleles are very common in the population, the probability to identify the *correct* victim is not very high. Identification is also difficult when the *Victim Model* is asked to cope with the case concerning the missing members in families 2, 8 and 9. In these cases the possibility to find the corresponding victims relies on considering more than one victim at time, so that, when the correct victims are introduced in the familial pedigree, they identify themselves exploiting the familial relationships.

On the opposite, using the *Family model*, the possibility to identify simultaneously groups of victims as the missing individuals in each family, allows to find the bodies corresponding to all the missing individuals in families 2 and 8. The limit of the *Family model* arises when it attempts to find the victims corresponding to the ninth family's missing persons. In this case, the correct victims receive a small identification probability since other two victims, V_2 and V_3, are in a stricter familial relationship, and have a higher probability to belong to the same familial group than victims V_{12} and V_{13}. Since, using the *Family model*, all the families are considered separately, the stricter structure of the missing persons in Family 2 with respect to Family 9 is not taken into account, assigning victims V_2 and V_3 to Family 2, where the stricter relationship is required. Finally, the *Complete model* provides a joint analysis of all the families and all the victims, and provides the correct answer in every circumstance.

Now analyze the *Intermediate stage* of the identification process. Here, victims labeled 8 and 14 are assumed not to be recovered and Families 4 and 7 have not claimed their relatives. Data are in Table (3) and results in Table

(4).

First consider the results concerning the missing individuals posed in parent-child relation with the claiming relatives. In Family 1, the available data and results are unchanged. The 7-th family does not appear since no claiming relatives are available. The 10-th family is still asking for identification, but the victim, corresponding to their missing relative, has not been recovered yet. In this case, since the victim is not available, the *Victim model* has nothing to say but the *Family model* and the *Complete model* assign, correctly, a very high probability to the event of no identification.

In the *Intermediate stage*, the identification of the missing person belonging to the fifth family is a difficult task also making use of the *Family* and the *Complete model*. This happens because the very common profile of the sixth victim provides support to the hypotheses to be one of the missing persons whose corresponding victims were not recovered yet.

Finally, consider the *Early stage* of the identification process, represented in table (5). Here, the 3-rd family does not claim its missing relative so that no result is available. Also, in the fifth family, the searched sibling is not among the victims as well as one sibling in each of the families 5, 8, 9, 10. Looking at (6), the unavailability of the victim in the fifth family is correctly detected by the *Family model* and the *Complete model* which assign a high probability to the no-identification event. A decay of the identification performances for the three missing persons in Family 8 happens because one of the victims has not been recovered, so that the identification of the right triplet is impossible. As a further consequence of the lack of information, both the Families 2 and 9 have now two victims recovered, without other familial evidence, so that a confusion between the two cases arises.

6. Conclusions

In this paper we have proposed a new model to identify the victim of a MFI. The starting point is the representation of an identification hypothesis comprising all the possible ways the recovered victims can be identified among the claimed missing persons. Then, inference is derived conditionally to all the evidence concerning the claiming families, the ethnicity of the missing persons and the genetic profiles of the recovered victims.

Conditional independence relations are of paramount importance to derive inference as it happens for graphical models which, unfortunately, cannot be directly employed given the very high dimensions of some conditional probability tables necessary to the analysis. Inference is analytically derived without resorting to standard propagation algorithms.

The identification of the victims of a MFI making use of DNA evidence is a task whose level of difficulty varies according to the available familial information and the sources of uncertainty we want to take into account.

Starting from familial information, if there is at least one direct lineage between the observed relatives and only one missing person is claimed by each family, all the identification models we discussed in this paper appear to be very reliable.

If, instead, the search is among relatives indirectly related or more missing persons are searched in the same family, the *Family* and the *Complete Model* fit the problem much better than the *Victim Model*.

Identifying more than one familial group, whose familial relationships is the only available evidence, is the most difficult task which can be safely accomplished only by the *Complete Model*, being some conditions satisfied. The most important is that of victims are recovered. Alternatively some further characteristics must be introduced to differentiate families, the most obvious being the gender of the missing persons. Another possibility is the introduction of some Y chromosome STR loci, even if this evidence can be used only if a paternal lineage is established between an observed member of a family and the missing person.

Finally two other sources of uncertainty, not included here but of some importance, must be mentioned.

First, some possible uncorrect specification of the kinship relations could be intentionally or unintentionally provided. A probabilistic treatment of such form of uncertainty is possible and outlined by Baio & Corradi (2007).

Then, a more detailed representation of the segregation process, through a convincing mutation model among those considered by Vicard et al. (2007) is attractive but some computational problems arise, related to the impossibility to restrict the analysis in each family to the observed alleles by means of a simple but computationally essential procedure, named recoding by Lauritzen & Sheehan (2003).

Researches on these fields have reached some encouraging results and will be provided in a forthcoming paper.

References

BAIO, G. & CORRADI, F. (2007). Handling Manipulated evidence. *Forensic Sci. Int.* 169, 181-7.

BRENNER, C. H. (1997). Symbolic Kniship Program. *Genetics* 145, 533-42.

Brenner, C. H. & Weir, B. (1997). Issues and strategies in the DNA identification of Word Trade Center victims. *Theoretical Population Biology* 63, 173-8.

Brenner, C. H. (2006). Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities. *Forensic Sci. Int.* 157, 172-80.

Cash, H.D., Hoyle, J. W. & Sutton, A. J. (2003). Development under Extreme Conditions: Forensic Bioinformatics int the wake Trade Center Disaster. In *Pacific Symposium Biocomputing*, 638-53.

Cavallini, D. & Corradi F. (2006). Forensic Identification of relatives of individuals included in a database of DNA profiles.*Biometrika*, 93, 525-36.

Clayton, T., Whitaker, J. & Maguire, C. (1995). Identification of bodies from scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci.*Forensic Sci. Int.*, 76, 7-15.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). Probabilistic Nwtworks and Expert Systems. Springer-Verlag.

Dawid, A. P., Mortera J., Pascali, V. L. & Boxel, D. V. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scand. J. Statist.*, 29, 577-95.

Dawid, A. P., Mortera J., Vicard, P. (2007). Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International*, 169, 195-205.

Evett, I. W. & weir, B. S. (1998). Interpreting DNA Evidence.*Sinauer Associates.*

Lauritzen, S. L. & Sheehan, N. A. (2003). Graphical models for genetic analyses. *Statist. Sc. Int.*, 18, 489-514.

Mortera, J., Dawid, A. P. & Lauritzen, S. L. (2003). Probabilistic expert system for DNA mixture profiling. *Theor. Population Biol.*, 63, 191-205.

Vicard, P., Dawid, A. P. & Mortera, J. (2007). Estimating mutation rates from paternity casework. *Statist. Sc. Int.*, 63, 191-205.

Weir, B. S. (1996). Genetic data analysis. *Sinauer Associates.*

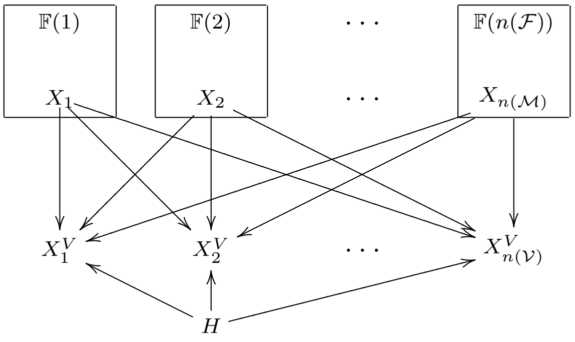

Figure 1: PES representation of the Complete Model for the MFI identification

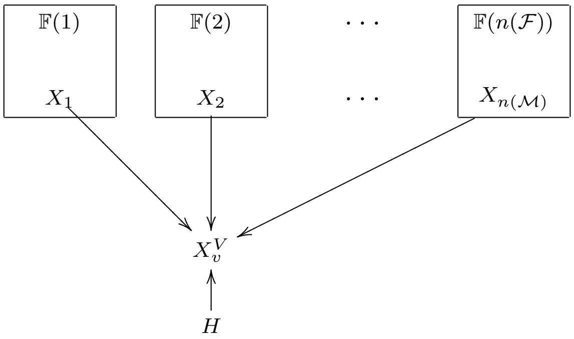

Figure 2: PES representation of the Complete Model for the MFI identification

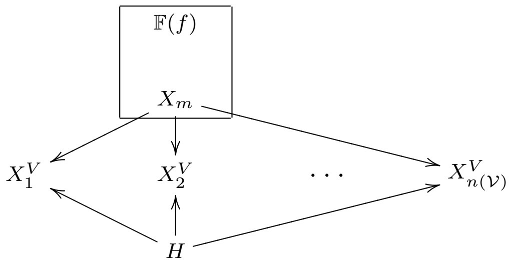

Figure 3: PES representation of the *Family model* for the MFI identification

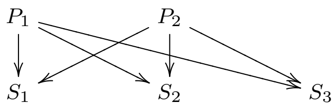

Figure 4: A generic family representation

Table 1: The *Final stage* available information

Family no.	Available relatives	Missing Persons	Victims
1	S_1, S_2	P_1	V_1
2	$---$	P_1, S_1	V_2,V_3
3	S_1	S_2	V_4
4	S_1	S_2	V_5
5	S_1	S_2	V_6
6	S_1	S_2	V_7
7	P_1, S_1	P_2	V_8
8	$---$	S_1, S_2, S_3	V_9,V_{10},V_{11}
9	$---$	S_1, S_2	V_{12},V_{13}
10	P_1	S_1	V_{14}

Table 2: The *Final stage*: posterior probabilities of correct identification

Families	1	2	2	3	4	5	6	7	8	8	8	9	9	10
Missing Persons	P_1	P_1	S_1	S_2	S_2	S_2	S_2	P_2	S_1	S_2	S_3	S_1	S_2	S_1
Victim Approach	.99	.28	.28	.99	.85	.02	.99	1	.42	.42	.42	.24	.24	1
Familial Approach	1	.99	.99	.99	.88	.96	.99	1	.99	.99	.99	.02	.02	1
Complete Approach	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 3: The *Intermediate stage* available information

Family no.	Available relatives	Missing Persons	Victims
1	S_1S_2	P_1	V_1
2	$---$	P_1, S_1	V_2, V_3
3	S_1	S_2	V_4
4	NA	NA	V_5
5	S_1	S_2	V_6
6	S_1	S_2	V_7
7	NA	NA	NA)
8	$---$	S_1, S_2, S_3	V_9,V_{10},V_{11}
9	$---$	S_1, S_2	V_{12},V_{13}
10	P_1	S_1	NA

Table 4: The *Intermediate stage* posterior probabilities of correct identification

Families	1	2	2	3	5	6	8	8	8	9	9	10
Missing Persons	P_1	P_1	S_1	S_2	S_2	S_2	S_1	S_2	S_3	S_1	S_2	S_1
Victim Model	.99	.22	.22	.99	.01	.99	.33	.33	.33	.22	.22	-
Familial Model	.99	.99	.99	.99	.06	.99	.99	.99	.99	.02	.02	.99
Complete Model	1	1	1	1	.11	1	1	1	1	1	1	.99

Table 5: The *Early stage* available information

Family no.	Available relatives	Missing Persons	Victims
1	$S_1 S_2$	P_1	V_1
2	$---$	P_1, S_1	V_2, V_3
3	NA	NA	V_4
4	NA	NA	V_5
5	S_1	S_2	NA
6	S_1	S_2	V_7
7	NA	NA	NA
8	$---$	S_1, S_2, S_3	V_9, V_{10}, NA
9	$---$	S_1, S_2	V_{12}, NA
10	NA	S_1	NA

Table 6: The *Early stage* posterior probabilities of correct identification

Families	1	2	2	5	6	8	8	8	9	9
Missing Persons	P_1	P_1	S_1	S_2	S_2	S_1	S_2	S_3	S_1	S_2
Victim Model	.99	.18	.18	-	.99	.18	.18	.18	.18	.18
Familial Model	.99	.99	.99	.99	.99	.04	.04	.04	.01	.01
Complete Model	.99	.22	.22	.99	.99	0.54	0.54	0.54	0.04	0.04