Mass Fatality Incident Identification based on nuclear DNA evidence

Fabio Corradi
Department of Statistics - University of Florence - Italy

Abstract

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 tragic 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. Finally, the proposal is tested through the simulation of a Mass Fatality Incident and the results are examined in details.

1 Introduction

Terrorists' attacks, natural calamities and transportation crashes have recently caused a relevant number of Mass Fatality Incidents (MFI), posing challenging identification problems to the authorities.

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, which is rarely available, 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 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 certain genetic population. The method derives from the solution of indirect identification problems like paternity cases. There, an individual is alleged to be in a certain position in a pedigree and no other alternative candidates are 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 and Weir (2003) by the consideration that the expected number of individuals, whose genetic profiles are compatible with the unobserved missing person's one, increases according to the population size and this is not negligible.

Actually the poor result obtained 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 not recovered 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, the families were still considered separately.

Our proposal consists in treating, simultaneously, all the victims and all the missing persons, evaluating an hypothesis random variable comprising all the possible identification conjectures. The proposed analysis is Bayesian since the identification hypothesis is unobservable and a prior probability is provided in a non-informative fashion, assuming that, before the DNA evidence become available, each victim has the same probability to be one of the missing persons. The analysis only requires familial groups not sharing recent ancestors with one another and a conservative estimate

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

of the number of victims.

2 Basic Ingredients

As a matter of notation calligraphic symbols indicates sets and $n(\cdot)$ 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, as usual, permutations, dispositions and combinations.

Let N the number of persons involved in a MFI. Assume N is exactly known, as it happens in the case of an aircraft incident and the passenger and the crew list is available or is guessed at a large conservative value. The aim is to identify the members of the set of recovered victims, $\mathcal{V}$, by means of the identity of the missing individuals, $\mathcal{M}$, claimed by the families, $\mathcal{F}$, who had some relatives disappeared in the MFI. To formalize the possibility that not all the victims have been recovered, we augment $\mathcal{V}$ by a ?, a generic no-recovered victim, so that $\mathcal{V}^* = \mathcal{V} \cup \{?\}$. Missing individuals in each family are detailed in $\mathcal{M}_f$, 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's members $\mathcal{U}$, connecting observed members. Also $\mathcal{M} = \cup_{f\in\mathcal{F}}\mathcal{M}_f$, $\mathcal{O} = \cup_{f\in\mathcal{F}}\mathcal{O}_f$, $\mathcal{U} = \cup_{f\in\mathcal{F}}\mathcal{U}_f$.

Let $H_m = v$, $v \in \mathcal{V}^*$, $m \in \mathcal{M}$, the hypothesis identifying the victim v as the m-th missing person. If this latter is considered in isolation, the identification random variable H_m can assume values in $\mathcal{V}^*$ without constraints. Instead, if more H_m are considered jointly, a multivariate random variable H must be defined so that its states take into account all the possible ways in which the missing individuals can identify the recovered victims since multiple assignments of the same victim to different missing persons are not allowed. Let H^t , a generic state of H called a configuration and conformed to the mentioned constraint:

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

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

$$n(\mathcal{H}) = D_{N,n(\mathcal{V})}. \tag{1}$$

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.

As a matter of notation $X^V = \{X^V_v : v \in \mathcal{V}\}$ refers to the recovered victims genotypes; $X^F = \{X^F_f : f \in \mathcal{F}\}$ deals with the families to which the missing persons belong and can be split into $X^F_f = \{X^M_f, X^O_f, X^U_f\}$, according to a partition of the family's into missing, observed and unobserved individuals being X^M , X^O and X^U the set of corresponding genotypes.

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, depending whether the individual's parents are included in the analysis, its 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 Mendel's law specifies the genotype' s probability of a child, c, given the genotypes of their parents, m and f. If $x_m = (i, j)$ and $x_m = (r, j)$, so that the set of the possible transmitted genotypes is $\mathcal{G} = \{\{i, r\}, \{i, s\}, \{j, r\}, \{j, s\}\}$ we have:

$$Pr(x_c|x_m, x_f) = 0.25 \sum_{g\in\mathcal{G}} I_{\{g\}}(x_c). \tag{2}$$

If mutations are taken into account, more sophisticated models are required, as in Dawid et al. (2007).

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

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

Inbreeding and co-ancestry characteristics in the populations can be included as in Evett and Weir (1998).

3 Model and inference

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

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

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 genotypes' random variables of the familial groups. This implies:

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

b) Familial groups are defined to 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 genotypes' random variables related to different families are independent:

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

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

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

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:

$$\begin{aligned} Pr(x^V, x^O|H^t) &= \sum_{X^M, X^U} Pr(x^V|X^M, H^t) \\ &\quad \cdot Pr(x^O, X^U, X^M). \end{aligned} \tag{8}$$

To evaluate (8), define as $\mathcal{M}_f^t = \{m \in \mathcal{M}_f : H_m^t \in \mathcal{M}_f : H_m^t \in \mathcal{V}\}$ the sets of missing persons in the families having victims assigned by a specified H^t and pose these families in the set $\mathcal{F}^t = \{f \in \mathcal{F} : \mathcal{M}_f^t \neq \emptyset\}$. Also let $\mathcal{X}_f^{M_t} = \{X_m : m \in \mathcal{M}_f^t\}$ 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:

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

where:

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

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):

$$\begin{aligned} &Pr(x_f^O, X_f^{M_t} = x_f^{V_t}) = \\ &\sum_{X_f^M, X_f^U} Pr(x_f^{V_t}|X_f^{M^t}, H^t) Pr(x_f^O, X_f^U, X_f^{M_t}), \end{aligned}$$

which is equivalent to a transfer of evidence from the victims to the assigned missing individuals. Finally:

$$\begin{aligned} &Pr(x^V, x^O|H^t) = \\ &\prod_{f\in\mathcal{F}^t} Pr(x_f^O, X_f^{M_t} = x_f^{V_t}) \prod_{f\in\mathcal{F}\setminus\mathcal{F}^t} Pr(x_f^O) \\ &= \prod_{f\in\mathcal{F}^t} \frac{Pr(x_f^O, X_f^{M_t} = x_f^{V_t})}{Pr(x_f^O)} \\ &= \prod_{f\in\mathcal{F}^t} Pr(X_f^{M_t} = x_f^{V_t}|x_f^O), \end{aligned} \tag{12}$$

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

As a result the likelihood is equal to the probability to observe each victim as if they would be the missing person designated by the H^t conditionally to the familial evidence.

3.1 Victims belonging to only one population

An intriguing formulation of the likelihood for H^t is possible if all missing persons belong to the same genetic population. In such case, (12) can be divided by the probability to observe all the recovered victims as belonging to the considered genetic population a constant not depending on H^t:

$$\begin{aligned} &Pr(x^V, x^O|H^t) \propto \\ &\prod_{f\in\mathcal{F}^t} \frac{Pr(X_f^{M_t} = x_f^{V_t}, x_f^O)}{Pr(x_f^O) \prod_{m\in\mathcal{M}_f^t : H_m^t = v} Pr(X_m^{M_t} = x_v^{V_t})} \\ &= \prod_{f\in\mathcal{F}^t} LR^t(f). \end{aligned} \tag{13}$$

This representation shows how the likelihood for H^t can be expressed by a number of likelihood ratios $LR^t(f)$, $f \in \mathcal{F}^t$, each one obtainable by a kinship analysis. More specifically, each $LR^t(f)$ 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 generic individuals belonging to the relevant genetic population. Both (12) and (13) allow

for computations at familial level, paving the way to parallel calculus strategies.

Furthermore, (13) points out that only families with $LR \neq 1$, are informative, and regards as useless those structurally unable to provide information to the hypothesis useless, since they always has $LR = 1$. In this latter category are families which have not claimed their missing person(s) yet. The case is formally represented by a missing individual searched by an empty family so that, if a victim is assigned, the corresponding LR is equal to one. This consideration has two important consequences. First, if some victims are not assigned for identification by a certain H^t, we can restrict the likelihood computation only to the potentially informative families and the associated missing individuals, respectively defined by:

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

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

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

It follows that, for a given H^t, not all the families contribute to the likelihood (13) but only those in the set $\mathcal{F}^t \cap \mathcal{F}^*$, so that the likelihood can be written as:

$$Pr(x^V, x^O|H^t) \propto \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} LR^t(f). \tag{14}$$

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

$$Pr(x^V, x^O|H^t) = Pr(x^V, x^O|H^s). \tag{15}$$

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 $t \neq s$, (15) 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 the members' likelihood.

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:

$$H^t_* = \{H^t_m : m \in \mathcal{M}^*\} \text{ where, } H^t_s =? \text{ or } \forall g \neq s \; H^t_g \neq H^t_s.$$

If a uniform prior is posed on the H^t', 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:

$$Pr(H^t_*|x^O, x^V) \propto D_{n(\mathcal{M}^+),n(\mathcal{V})-i^t} \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} LR^t(f). \tag{16}$$

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 $\mathcal{I} = \{max(0, n(\mathcal{V}) - n(\mathcal{M}^+)), \ldots, min(n(\mathcal{M}^*), n(\mathcal{V}))\}$. The number of possible equivalent classes for each i is $C_{n(\mathcal{V}),i} \cdot C_{n(\mathcal{M}^*),i} \cdot P_i$, and:

$$n(\mathcal{H}^*) = \sum_{i \in \mathcal{I}} C_{n(\mathcal{V}),i} \cdot C_{n(\mathcal{M}^*),i} \cdot P_i. \tag{17}$$

The saving in computational efforts can be evaluated case-by-case comparing (17) with (1).

A noticeable case arises if, in a familial group, the relationships are known but no genetic 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, $LR \neq 1$, since the probability to observe the victims, evaluated assuming the familial relationship, differs if the assumption of independence holds.

This makes it possible to identify victims also in these extreme circumstances.

3.2 Victims belonging to more populations

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 (13) now varies from one configuration to another, depending on which genetic population the missing persons, who have victims assigned, belong to.

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 (12), splitting the product into informative and non informative families:

$$Pr(x^V, x^O|H^t) \propto \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} \frac{Pr(x_f^O, X_f^{M_t} = x_f^{V_t})}{Pr(x_f^O)} \cdot \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^+} Pr(X_f^{M_t} = x_f^{V_t}). \quad (18)$$

If we multiply and divide (18) by the probability to observe the victims, arranged according to $\mathcal{F}^*$ and $\mathcal{F}^+$, we get the likelihood expression:

$$Pr(x^V, x^O|H^t) \propto \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^*} LR^t(f) \prod_{m \in \mathcal{M}_f^* : H_m^t = v} Pr(X_m^{M_t} = x_v^{V_t}) \cdot \prod_{f \in \mathcal{F}^t \cap \mathcal{F}^+} \prod_{m \in \mathcal{M}_f^+ : H_m^t = v} Pr(X_m^{M_t} = x_v^{V_t}), \quad (19)$$

where the likelihood ratios for the informative families in (16) still appear but the probability to observe the victims now depends on the population of the families to which they are assigned by H^t.

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 allotted, this probability varies according to the population the unclaimed missing individuals belong to. This ruins the idea of inferential equivalent classes but it is still convenient to express the likelihood for each H_*^t, by 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,

$$N_i^+ = N\pi_i - \sum_{f \in \mathcal{M}_f^*} \sum_{m \in f} I_{\{i\}}(G_m), \forall i \in \mathcal{K}, \quad (20)$$

and their total number,

$$N^+ = \sum_{i \in \mathcal{K}} N_i^+, \quad (21)$$

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:

$$Pr(G) = \prod_{m \in \mathcal{M}_f^+ : H_m^t = v} Pr(G_m|G_{-m} = g_{-m}) = \frac{\prod_{i=1}^k D_{N_i^+, n_i^t}}{D_{N^+, n_t}}, \quad (22)$$

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:

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

which represents the generalization of (16) to k populations.

3.3 Some computational remarks

To make inference on H, two computational issues must be efficiently addressed: the evaluation of the likelihood for each family with one or more victims assigned by a configuration, and how to get the states of the variable H.

The familial likelihood evaluation is performed, according to (12) by a single propagation in an Allele Bayesian Network (Lauritzen and Sheehan (2003)), which efficiently derives, for each locus and family, the probability distribution of the missing persons, conditionally to x_f^O. Then it is computationally inexpensive to evaluate the probability of every possible set of recovered victims attempting to be identified as the missing person of a certain family. If the segregation model (2) is used, some of the victims may have probability zero to be identified as some of the claimed

missing persons: this could happen if the donors and the missing person are on a direct lineage or if three or more of the missing person's siblings provide their genetic profiles. As a result we get, for each family, a list of victims to be assigned to the H^t configurations for which (8) results strictly greater than zero. Following this consideration a simple but efficient procedure to produce the relevant H states has been implemented.

Consider the families and the associated lists of victims potentially identifiable as their missing members. Assume initially that just one victim is identified ($i = 1$). The H states' list is straightforward and the number of states is equal to the sum of the different victims in each family's lists.

To find the configuration with two or more victims assigned, the following pseudo code can be used.

1) $i = i + 1$;

2) consider a configuration found in the previous victim assignment procedure and remove the victim(s) already assigned from the lists of victims' candidates in the other families;

3) produce as many as possible new states of H with an additional victim assigned and taken from the families' lists of victims obtained in the previous step, avoiding multiple victims imputations among and inside the families;

4) repeat the process from 2) until all the configurations found in the victim assignment procedure with $i - 1$ victim are considered;

5) repeat the process from 1) until the maximum number of assignable victims is reached, i.e. when $i + 1 > min(n(\mathcal{M}^*), n(V))$.

4 A simulation-based example

In this section we display the results of an identification process carried on 14 individuals belonging to 10 families and disappeared in a simulated MFI.

To make it possible to display the results in details, the example proposes a small number of individuals and families, but many of the difficulties which typically arise in the field are included. Individuals in the families are classified as parents (P) and siblings (S).

Data are simulated to have the opportunity to check the results, since the victims' identity is known. Ancestors' genotypes are simulated from a genetic population according to (3); missing persons' profiles are sampled and posed in the victims' data set. To produce the result we made use of a Bayesian network.

A graph representation for a family with two parents and three siblings is in Fig. 1.

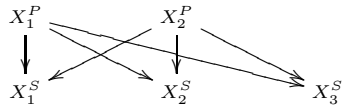

Figure 1: Graph representation of a two parents - three siblings family

Information concerning the donors, the missing persons and the corresponding victims in the example are detailed in Tab. 1 for each family.

Table 1: Set of individuals relevant for the analysis: Complete information case

f	$\mathcal{O}_f$	$\mathcal{M}_f$	$\mathcal{V}_f$
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}

In the example, families 1, 7 and 10, search for only one missing person posed in direct lineage with the claiming relatives; in families 3, 4, 5, and 6 one sibling is looking for another sibling; in the remaining three families 2, 8 and 9, the search considers more than one missing person in each family, but no donor is available. Starting from uninformative prior probabilities on every possible configuration H, the exercise evaluates the identification probabilities comparing results obtained by the proposed model, called the *Full Model*, and those obtained by two simpler models, the *Victim* and the *Family Models*.

The *Full Model* is represented in Fig. 2 by a graph, embedding all the conditional independence assertions implied in (6) and (7) and all the relations among missing individuals, their relatives, the victims and the identification hypothesis. In the boxes, the families' members and their relations, made explicit in Fig. 1, are hidden, and only the relevant missing individuals' unobserved genotypes are considered.

The *Victim model*, Fig. (3), restricts the victim set to one victim only, attempting the identification through a search among the missing persons. This model essentially represents the standard practice consisting in

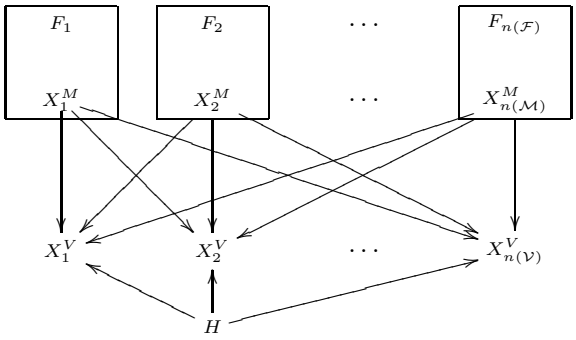

Figure 2: Graph representation of the *Full Model*

the attempt to identify a victim at a time even if all the possible missing individuals are contemporary considered as in Cavallini and Corradi (2006).

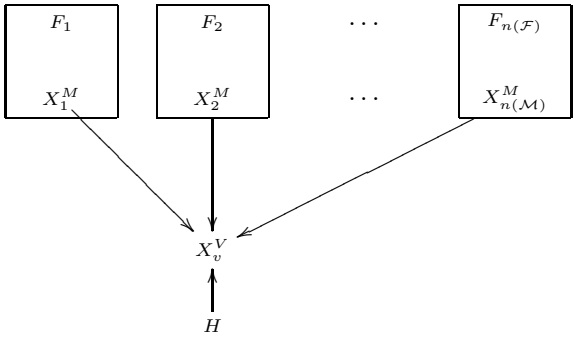

Figure 3: Graph representation of the *Victim Model*

The *Family model*, Fig. (4), considers a family trying to find, among the victims, their lost members. This approach is similar to the model proposed by Brenner (2006).

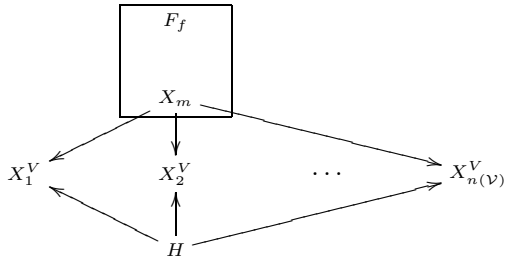

Figure 4: Graph representation of the *Family model*

In Tab. 2, for each missing person, we provide the posterior probabilities that each model assigns to the victim who actually is the missing individual.

Note that, every time the observed relatives and the claimed missing persons are in a direct lineage, all the models provide a very high posterior probability of

Table 2: Probabilities of correct identification - Complete information case

		Models		
Family		*Full*	*Family*	*Victim*
1	P_1	1	0.99	0.99
2	P_1	1	0.99	0.28
	S_1	1	0.99	0.28
3	S_2	1	0.99	0.99
4	P_1	1	0.88	0.85
5	S_2	1	0.96	0.02
6	S_2	1	0.99	0.99
7	P_2	1	1	1
8	S_1	1	0.99	0.42
	S_2	1	0.99	0.42
	S_3	1	0.99	0.42
9	S_1	1	0.02	0.24
	S_2	1	0.02	0.24
10	S_1	1	1	1

correct identification. This happens because the first Mendel law produces a large number of exclusions, assigning zero probability to all the victims incompatible with the families' donors. The same is not true if we consider families 3, 4, 5, 6 where a member of the family is looking for a sibling: For instance, the *Victim* Model does not succeed in the identification of V_6 as the missing person in the 5-th family: this happens because V_6 has a very common genetic profile, so that only consideration of the other victims allows to get a high probability of correct identification. Identification is also difficult for the *Victim Model* when only the family structure is known as it happens for families 2, 8 and 9. Here the possibility to find the corresponding victims relies on considering more than one victim at a 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 missing persons in the ninth family. In this case, V_{12} and V_{13}, actually belonging to the family, receive a small identification probability since other two victims, V_2 and V_3, are more strictly related. In this case only the *Full* model, jointly considering all the families, provides the correct answer.

To simulate the possibility that some pieces of information are not available yet, as it happens at an early stage of the identification process, we hide some of recovered victims and of the claimed missing persons as

it is shown in Tab. 3. The results obtained from the competition of the three models are in Tab. 4.

Table 3: Set of individuals relevant for the analysis: Incomplete information case

f	$\mathcal{O}_f$	$\mathcal{M}_f$	$\mathcal{V}$
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: Probabilities of correct identification - Incomplete information case

		Models		
Family		*Full*	*Family*	*Victim*
1	P_1	1	0.99	0.99
2	P_1	1	0.99	0.22
	S_1	1	0.99	0.22
3	S_2	1	0.99	0.99
5	S_2	0.11	0.06	0.01
6	S_2	1	0.99	0.99
8	S_1	1	0.99	0.33
	S_2	1	0.99	0.33
	S_3	1	0.99	0.33
9	S_1	1	0.02	0.22
	S_2	1	0.02	0.22
10	S_1	0.99	0.99	–

When the information is incomplete, finding the victim who is the missing person belonging to the fifth family becomes difficult also making use of the *Family* and *Full Models*. This happens because the very common profile of V_6 provides support to the hypotheses he/she is one of the missing persons whose corresponding victim has not been not recovered yet.

5 Conclusions

In this paper we proposed a new model to identify victims of a Mass Fatality Incident. 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 genetic evidence concerning the claiming families, the ethnicity of the missing persons and the genetic profiles of the recovered victims. 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 to be taken into account. Identifying more than one familial group, whose familial relationships are the only available evidence, is the most difficult task which can be safely accomplished only if a large fraction of their victims are recovered.

To make as short as possible the list of victims eligible for the identification of each missing person, some further characteristics of these latter need to be introduced, the most obvious is the missing persons' gender but it proves useful to consider some Y chromosome STR loci and Mt-DNA fraction used for identification. This strategy obviously does not apply if familial donors and the missing person are not on the paternal or maternal lineage or if the familial evidence is not available at all.

References

Brenner, C. (1997). Symbolic Kinship Program. *Genetics 145*, 533–542.

Brenner, C. (2006). Some Mathematical problems in the DNA Identification of Victims in the 2004 Tsunami and similar Mass Fatalities. *Forensic Sci. Int. 157*, 172–180.

Brenner, C. H. and B. Weir (2003). Issues and Strategies in the DNA Identification of Word Trade Center Victims. *Theoretical Population Biology 63*, 17–38.

Cash, H., J. W. Hoyle, and A. J. Sutton (2003). Development under Extreme Conditions: Forensic Bioinformatics int the Wake Trade Center Disaster. In *Pacific Symposium Biocomputing*, pp. 638–53.

Cavallini, D. and F. Corradi (2006). Forensic Identification of Relatives of Individuals Included in a Database of DNA Profiles. *Biometrika 93*, 525–36.

Clayton, T., J. Whitaker, and C. Maguire (1995). Identification of Bodies from Scene of Mass Disaster Using Amplification of Short Tandem Repeat (STR) Loci. *Forensic Sci. Int. 76*, 7–15.

Dawid, A., J. Mortera, and P. Vicard (2007). Object-Oriented Bayesian Networks for Complex Forensic DNA Profiling Problems. *Forensic Science International 169*, 195–205.

Evett, I. and B. Weir (1998). *Interpreting DNA evidence*. Sinauer Associates, Sunderland.

Lauritzen, S. and N. Sheehan (2003). Graphical Models for Genetic Analyses. *Statistical Science 18*, 489–514.

Weir, B. (1996). *Genetic Data Analysis*. Sinauer Associates, Sunderland.