

A SIMULATION-BASED MODEL OF ABDUCTION

Gildas Morvan
LGI2A - EA 3926
Université d'Artois
Technoparc Futura
F-62400 Béthune
email: gildas.morvan
@fsa.univ-artois.fr

Daniel Dupont
Département Ingénierie
et sciences du vivant
HEI, 13 rue Toul
F-59800 Lille

Philippe Kubiak
LAGIS - UMR CNRS 8146
École Centrale de Lille
Cité Scientifique - BP 48
F-59651 Villeneuve d'Ascq
email: philippe.kubiak@ig2i.fr

KEYWORDS

simulation-based reasoning, abduction, forensic entomology

ABSTRACT

Abduction, or Inference to the Best Explanation (IBE), is a reasoning process that generates possible explanations from a set of "surprising" observations. In this paper, a simulation-based model of abduction is introduced. This model is then implemented to develop a decision support system in the field of Forensic Entomology to help forensic scientists solve complex cases.

INTRODUCTION

Abduction, or Inference to the Best Explanation (IBE), is a reasoning process that generates possible explanations from a set of "surprising" observations. It has been widely studied in Philosophy and later on, in Artificial Intelligence (AI). The literature on this subject is so vast that a whole book would not be enough to undertake a complete survey. Therefore, only the key aspects of some of the abduction models developed in AI, set cover and logic based approaches, are briefly introduced. However readers may refer to Paul (1993) for a more comprehensive presentation.

Set cover based approaches define an abductive framework as a triplet $\langle \Phi, \Omega, e \rangle$ where Φ is a set of hypotheses, Ω is a set of manifestations, and e is a mapping from 2^Φ to 2^Ω . The abductive problem is then defined as follows: Let $\Omega^* \subseteq \Omega$ be a set of observed manifestations. The problem is to determine a minimal set $\Phi^* \subseteq \Phi$ such as $e(\Phi^*) = \Omega^*$.

Logic based approaches assume that domain knowledge is encapsulated in a theory \mathcal{T} defined over a language \mathcal{L} . Let \mathcal{A} , a set of sentences of \mathcal{L} , be a set of hypotheses. Let the sentence ω be a *surprising observation* (i.e., $\mathcal{T} \not\vdash \omega$). ϕ is an explanation of ω iff $\mathcal{T} \cup \phi \vdash \omega$, $\mathcal{T} \cup \phi$ is consistent and $\phi \in \mathcal{A}$.

Existing abduction AI models rely on knowledge models structured a specific language, defining truth values for propositions or relations between facts. However, the

knowledge about a given system can be encapsulated in many types of models, e.g., a computational model. The abduction model presented in this paper does not assume an *a priori* knowledge of causality relations nor a specific syntax for the knowledge base, but assumes that: (1) the set of possible explanations (hypotheses) can be defined as a *metric space* — i.e., distance measures can be computed between all the hypotheses; (2) it exists a deductive (predictive) model able, e.g., by simulating the causal history of the system, to compute a *coherence measure* between a hypothesis and a set of observations. Then, this model is applied to develop a decision support system in the field of Forensic Entomology, using agent-based simulations. Experimental results show that using this model to solve complex cases increases the efficiency and the precision of the results.

A SIMULATION-BASED MODEL OF ABDUCTION

Definitions

Let Ω be a set of observations and Φ a set of hypotheses — i.e., possible explanations — for Ω . We assume that a metric space (Φ, d^Φ) can be defined. A coherence measure between an hypothesis ϕ of Φ and Ω is computed by a deductive (predictive) model $m : \Phi \times \Omega \rightarrow [0, 1]$. This coherence measure is denoted $c_{m,\phi,\Omega}$.

Definition 1 A hypothesis is called *validated* iff its coherence measure is known. Let $\Phi^v \subseteq \Phi$ be the set of validated hypotheses and $\mathcal{R}_{m,\Phi^v,\Omega} : \Phi \times [0, 1]$ be the solution set computed by a model m :

$$\mathcal{R}_{m,\Phi^v,\Omega} = \bigcup_{\forall \phi \in \Phi^v} (\phi, c_{m,\phi,\Omega}). \quad (1)$$

Similarly, a metric space $(\mathcal{R}_{m,\Phi^v,\Omega}, \widehat{d}^{\mathcal{R}})$ is defined. An estimated coherence measure $\widehat{c}_{m,\phi,\Omega}$ is associated to all ϕ in $\Phi - \Phi^v$. This measure is estimated from $\mathcal{R}_{m,\Phi^v,\Omega}$ by interpolation. The resulting estimated solution set is defined as follows:

$$\widehat{\mathcal{R}}_{m,\Phi,\Omega} = \bigcup_{\forall \phi \in \Phi} \begin{cases} (\phi, c_{m,\phi,\Omega}) & \text{if } \phi \in \Phi^v \\ (\phi, \widehat{c}_{m,\phi,\Omega}) & \text{otherwise.} \end{cases} \quad (2)$$

Goal g	Definition
<i>pr.</i>	Determine the most probable hypotheses
<i>el.</i>	Eliminate unprobable hypotheses
<i>prs.</i>	Determine $c_{m,\phi,\Omega}$ for probable hypotheses
<i>id.</i>	Determine $c_{m,\phi,\Omega}$ for every hypotheses

Table 1: Goals of an abductive task

For convenience, in the rest of the paper, metric spaces will be denoted as their respective sets. During the abductive process, hypotheses to be validated are chosen iteratively by a heuristic

$$\gamma : \widehat{\mathcal{R}_{m,\Phi,\Omega}} \rightarrow \Phi - \Phi^v. \quad (3)$$

Abduction goals

It is generally accepted that abduction targets the best explanation of a given set of observations. However, in the context of a decision support system, this definition should be expanded. In the table 1, four different goals for an abductive task are defined. Thus, in its most general definition, abduction can be seen as an identification task (goal *id.*).

Quality measures can be defined to determine how goals have been achieved. A quality measure, $q_{m,g}$, based on the root mean square deviation (RMSD) between $\mathcal{R}_{m,\Phi,\Omega}$ and $\widehat{\mathcal{R}_{m,\Phi^v,\Omega}}$ is defined for each goal g of the table 1. *E.g.*, for the goal *id.* the quality measure is defined as follows:

$$q_{m,id.} = \sqrt{\frac{\sum_{\forall \phi \in \mathcal{R}_{m,\Phi,\Omega}} (c_{m,\phi,\Omega} - c_{m,\phi,\Omega})^2}{|\Phi|}}. \quad (4)$$

For the other goals, subsets of $\mathcal{R}_{m,\Phi,\Omega}$ and $\widehat{\mathcal{R}_{m,\Phi^v,\Omega}}$, meeting their defining condition — *e.g.*, for the goal *prs.*, subsets are defined such as, for any ϕ in $\mathcal{R}_{m,\Phi,\Omega}$, $c_{m,\phi,\Omega} > 0$ — are used.

Knowing the result set is mandatory to use these quality measures. Thus, they are useful to evaluate an implementation of the algorithm — and especially the heuristic γ —, not a particular solution.

Resolution algorithm

The algorithm 1 computes an estimated solution set $\widehat{\mathcal{R}_{m,\Phi^v,\Omega}}$ using a deductive model m . One could argue that it is not an abductive algorithm as it does not return the best explanations for Ω . Nevertheless, as it explains possible causal histories of Ω , it is, in some ways, similar to the abductive framework developed in Lipton (2004; ch. 4). Moreover, the solution set can be used as an input for post-processing algorithms, *e.g.* to merge solution sets obtained with different deductive models, and then determine the best explanation. Halting of the algorithm 1 depends on a condition C . In the implementation presented in the next section, the following

condition is used:

$$\sum_{\forall r \in \mathcal{R}_{m,\Phi^v,\Omega}} cost(r) \leq maxCost, \quad (5)$$

where $cost$ represents the cost necessary to select and validate an hypothesis and $maxCost$ is the maximal cost allowed by the user to compute the solution. Of course the algorithm halts if all the ϕ in Φ have been validated. Using a condition of the type

$$c_{m,\phi,\Omega} \geq \mu, \quad (6)$$

where μ is a validity threshold, the algorithm would be a form of the hypothetico-deductive model (Lipton 2004; ch. 4).

Algorithm 1 Inquisitive abduction algorithm

Input: A set of observations Ω

Input: A halting condition C

Input: A heuristic γ

Output: A solution set $\widehat{\mathcal{R}_{m,\Phi^v,\Omega}}$

- 1: Determine Φ
 - 2: $\Phi^v \leftarrow \emptyset$
 - 3: $\mathcal{R}_{m,\Phi^v,\Omega} \leftarrow \emptyset$
 - 4: compute $\widehat{\mathcal{R}_{m,\Phi,\Omega}}$
 - 5: **while not** (C **or** ($\Phi = \Phi^v$)) **do**
 - 6: $\phi \leftarrow \gamma(\widehat{\mathcal{R}_{m,\Phi,\Omega}})$
 - 7: $c_{m,\phi,\Omega} \leftarrow m(\phi, \Omega)$
 - 8: $\Phi^v \leftarrow \Phi^v \cup \phi$
 - 9: $\mathcal{R}_{m,\Phi^v,\Omega} \leftarrow \mathcal{R}_{m,\Phi^v,\Omega} \cup (\phi, c_{m,\phi,\Omega})$
 - 10: compute $\widehat{\mathcal{R}_{m,\Phi,\Omega}}$
 - 11: **end while**
 - 12: **return** $\widehat{\mathcal{R}_{m,\Phi^v,\Omega}}$
-

In the worst case, *i.e.*, the halting condition C stays false during the entire execution, this algorithm needs $|\Phi|$ steps to halt and return $\mathcal{R}_{m,\Phi,\Omega}$, *i.e.*, the exact solution set. However, the complexity of the algorithm depends mainly on the complexity of the heuristic and on the solution set estimation algorithm. This algorithm is very general; thus, it does not refer to the goal of the abductive task; this is handled by the heuristic γ . In the next section, elements for constructing such heuristics are presented. We focus on the goals *el.*, *prs.* and *id.*: for the goal *pr.*, classical optimisation heuristics can be used efficiently. Moreover, we target cases where the hypothesis set is not very big but the cost of validation of a hypothesis is important.

HEURISTIC DEFINITIONS

Definition of the criteria

Many criteria that help an agent to choose the best hypothesis to validate have been identified in the literature. Among all of these, simplicity seems to have been

the most used by computer scientists. Simplicity has been interpreted as logical simplicity. But it seems obvious that for a cognitive agent, simplicity is a little more complex (Aliseda-Llera 1998, Paul 1993). In the model presented in this paper, simplicity is defined as follows:

Definition 2 A hypothesis ϕ is simple if it exists a retrodictive model m^{-1} such as $m^{-1}(\Omega) = \phi$. Moreover, if Φ is bounded, ϕ should be considered as simple if $\phi = \inf(\Phi)$ or $\phi = \sup(\Phi)$.

The simplicity criterion is used to preprocess the problem. Thus, all the hypotheses defined as simple will be primarily chosen and validated.

Other criteria such as cost or utility have been quoted and should be used to handle the choice of a hypothesis in an abductive task (Peirce 1931, McGrew 2003). The cost criterion can be defined informally as follows: the selected hypothesis should be, *ceteris paribus*, the one that minimises the cost of validation. This criterion should be used if it is possible to estimate empirically a cost function of the model m , $cost_m : \Phi \rightarrow \mathbb{R}$.

Let $u_c : \Phi \rightarrow [0, 1]$ be a utility function of the cost criterion. For any ϕ in Φ

$$u_c(\phi) = 1 - \frac{cost_m(\phi)}{max(cost_m)}. \quad (7)$$

The utility criterion can be interpreted in different ways. In this approach, it could be defined as follows: the selected hypothesis should be, *ceteris paribus*, the one which maximises the knowledge of the agent. It means that the chosen hypothesis should be the one which best improves the quality of the solution — relatively to a quality measure. Of course in a real world problem, such an hypothesis as well as its effect on the quality measure cannot be determined with certainty. However, the metric of Φ or $\mathcal{R}_{m, \Phi^v, \Omega}$ can be used to estimate a utility function for this criterion. First, consecutivity is defined:

Definition 3 Let (E, d) be a metric space. Two elements e_i and e_j of E are consecutive in (E, d) iff it does not exist an element e of E such as

$$d(e, e_i) < d(e_i, e_j) \text{ and } d(e, e_j) < d(e_i, e_j). \quad (8)$$

The notion of maximal distance between two consecutive elements in a metric space is then defined as follows:

Definition 4 Let (E, d) be a metric space and $d_{max}((E, d))$ the maximal distance between two consecutive elements in (E, d) . For any consecutive elements e_i and e_j of E , $d_{max}((E, d)) = d(e_i, e_j)$ iff it does not exist two consecutive elements e_k and e_l in (E, d) such as $d(e_k, e_l) > d(e_i, e_j)$.

Two different utility functions of the utility criterion can then be defined using the metrics of Φ — equation 9 —

and $\mathcal{R}_{m, \Phi, \Omega}$ — equation 10. Let $u_u^\Phi, u_u^\mathcal{R} : \Phi \rightarrow [0, 1]$ be two utility functions of the utility criterion. For any ϕ of Φ

$$u_u^\Phi(\phi) = \begin{cases} 0 & \text{if } \phi \in \Phi^v \\ \frac{d^\Phi(\phi, \phi_1)}{d_{max}^\Phi(\Phi^v)} & \text{otherwise,} \end{cases} \quad (9)$$

where the hypothesis $\phi_1 \neq \phi$ is the closest to ϕ in $\Phi^v \cup \{\phi\}$, i.e., there is no $\phi_2 \neq \phi$ in $\Phi^v \cup \{\phi\}$ such as $d^\Phi(\phi, \phi_2) < d^\Phi(\phi, \phi_1)$.

For any $r = (\phi, c_{m, \phi, \Omega})$ of $\mathcal{R}_{m, \Phi, \Omega}$ — $c_{m, \phi, \Omega}$ may be unknown

$$u_u^\mathcal{R}(\phi) = \begin{cases} 0 & \text{if } \phi \in \Phi^v \\ \frac{d^\mathcal{R}(r, r_1)}{d_{max}^\mathcal{R}(\mathcal{R}_{m, \Phi^v, \Omega})} & \text{otherwise,} \end{cases} \quad (10)$$

where the result $r_1 \neq r$ is the closest to r in $\mathcal{R}_{m, \Phi^v, \Omega} \cup \{r\}$, i.e., there is no $r_2 \neq r$ in $\mathcal{R}_{m, \Phi^v, \Omega} \cup \{r\}$ such as $d^\mathcal{R}(r, r_2) < d^\mathcal{R}(r, r_1)$.

For any ϕ of Φ , if ϕ belongs to Φ^v , $u_u^\Phi(\phi) = u_u^\mathcal{R}(\phi) = 0$. This is a desired property as it is useless to validate a hypothesis twice with the same deductive model.

Heuristic definitions

An aggregation operator is used to determine the global utility functions u^Φ and $u^\mathcal{R}$ — using respectively u_u^Φ and $u_u^\mathcal{R}$ — of a hypothesis. As an example, weighted sum is used here. However any aggregation operator can also be used.

Let $u^\Phi, u^\mathcal{R} : \Phi \rightarrow [0, 1]$ be two functions representing the utility of any hypothesis of Φ . For any ϕ of Φ

$$u^\Phi(\phi) = \alpha_\Phi \cdot u_c(\phi) + \beta_\Phi \cdot u_u^\Phi(\phi), \quad (11)$$

with $\alpha_\Phi + \beta_\Phi = 1$.

For any $r = (\phi, c_{m, \phi, \Omega})$ of $\mathcal{R}_{m, \Phi, \Omega}$

$$u^\mathcal{R}(\phi) = \alpha_\mathcal{R} \cdot u_c(\phi) + \beta_\mathcal{R} \cdot u_u^\mathcal{R}(\phi), \quad (12)$$

with $\alpha_\mathcal{R} + \beta_\mathcal{R} = 1$.

Two heuristics γ_Φ and $\gamma_\mathcal{R}$ can now be defined.

Definition 5 Let γ_Φ (resp. $\gamma_\mathcal{R}$) : $\mathcal{R}_{m, \Phi, \Omega} \rightarrow \Phi - \Phi^v$ be a function that chooses an hypothesis to be validated.

For any $r = (\phi, c_{m, \phi, \Omega})$ of $\widehat{\mathcal{R}_{m, \Phi, \Omega}}$, $\gamma_\Phi(\widehat{\mathcal{R}_{m, \Phi, \Omega}})$ (resp. $\gamma_\mathcal{R}(\widehat{\mathcal{R}_{m, \Phi, \Omega}})$) = r iff it does not exist a $r' = (\phi', c'_{m, \phi', \Omega})$ in $\widehat{\mathcal{R}_{m, \Phi, \Omega}}$ such as $u^\Phi(\phi') > u^\Phi(\phi)$ (resp. $u^\mathcal{R}(\phi') > u^\mathcal{R}(\phi)$) and ϕ does not belong to Φ^v .

The efficiency of these heuristics is evaluated on a real world problem presented in the following section. The goals *el. et id.* are considered.

APPLICATION TO FORENSIC ENTOMOLOGY

Introduction to Forensic Entomology

Forensic entomology is widely used in criminal investigations to determine post-mortem intervals (PMI) from

the insects found on a cadaver. A PMI is usually estimated by experts using retrodictive models. These models are very easy to use but they do not take into account the ecosystemic context; thus, estimations performed using these methods are often overestimated and not as precise as they could be. Modern PMI estimation methods are based on insect development models. These models consider that insect development speed is temperature-dependant (Stinner et al. 1974). It is given as a function f of the temperature T varying in the time t . When a cadaver is discovered, investigators take insect samples from the body. Entomologists determine the species and the accumulated rates of development, denoted Δa , of the oldest individuals. Then, for each one of them, the laying time t_1 (generally close to the time of death) can be calculated from the following equation

$$\Delta a = \int_{t_1}^{t_2} f(T(t))dt, \quad (13)$$

where t_2 represents the time of the cadaver discovery. Data from the nearest meteorological station are usually used in order to estimate $T(t)$. However, it is problematic for many reasons. First, one can notice that cadavers are rarely found at the foot of a meteorological station. Furthermore, corpse thermal inertia is important, especially in the first hours after death. Finally the heat generated by larva aggregates can raise the temperature locally up to 20 °C. Thus, in many cases, entomological expertises results are inaccurate and given with an important margin of error.

A decision support system for Forensic Entomology

To handle these issues, a predictive agent-based model of insect development and cadaver decomposition in a complex ecosystem has been developed. This model is used to determine if a hypothesis — a possible time of death — is coherent with the observations available on the ecosystem of the crime scene and the entomofauna found on the victim. More information about this model and the validation process can be found in Morvan et al. (2007).

As the model is stochastic, it is necessary to run a large number — about 100 in most of the cases — of simulations to compute a coherence measure statistically significant. For any ϕ of Φ^v

$$c_{m,\phi,\Omega} = \frac{s_{m,\phi}^v}{s_{m,\phi}^t}, \quad (14)$$

where $s_{m,\phi}^v$ represents the number of valid simulations and $s_{m,\phi}^t$ the total number of simulations. Possible times of death are obtained from police investigators. Let ϕ_0 be the last time the victim was seen alive and ϕ_n the time she was found. The set of possible times of death

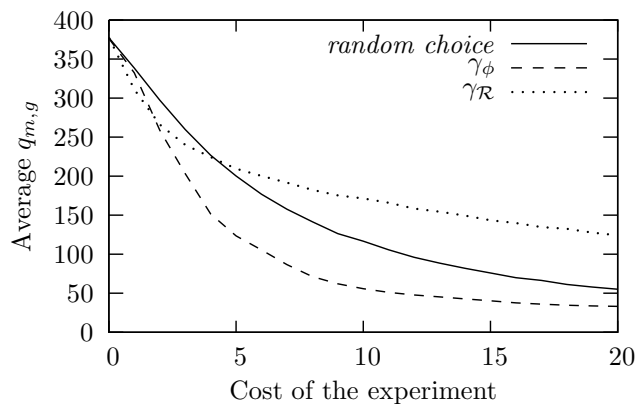


Figure 1: Quality of the heuristics is evaluated for the goal id . and compared to a random choice of hypotheses.

is discretised hour by hour: $\Phi = \{\phi_0, \dots, \phi_n\}$; a finer discretisation would not be mandatory unless more precise climatic data can be obtained. The metric spaces (Φ, d) and $(\mathcal{R}_{m,\Phi,\Omega}, d)$ are defined with the euclidean distance d . Simple hypotheses are ϕ_0 and ϕ_n and any time of death determined with a "classical" method — cf. equation 13. Estimation of the result set is performed using linear interpolation. In the next section, some experimental results are presented.

Experimental results

To evaluate the quality of the heuristics, "realistic" solution sets — *i.e.*, with the same fuzzy Gaussian shape than real solution sets — are randomly generated with $|\Phi| = 200$. The cost function of the agent-based model m has been approximated using the following function: for any ϕ_i of Φ

$$cost_m(\phi_i) = \lambda \cdot \left(1 - \frac{i}{|\Phi| - 1}\right), \quad (15)$$

where λ represents the validation cost of the hypothesis ϕ_0 .

Here we assume that $\lambda = 1$. Thus, 50.25 would be necessary to validate all the hypotheses of Φ . The maximal cost allowed to compute the solution is arbitrary fixed to 20. 10000 result sets have been generated to determine the best affectation for α_{ϕ} and $\alpha_{\mathcal{R}}$. Results show that whatever the goal is $\alpha_{\phi} \simeq \alpha_{\mathcal{R}} \simeq 0.001$. An interesting feature of this result is that the quality of the solution is very bad if the cost criterion is not used — *i.e.*, if $\alpha_{\phi} = \alpha_{\mathcal{R}} = 0$ — and best for very small values of α_{ϕ} and $\alpha_{\mathcal{R}}$. Results presented in the figures 1 and 2 show that γ_{ϕ} and $\gamma_{\mathcal{R}}$ perform well respectively for the goals id . and el .

Many different development models of Diptera can be found in the literature. It has been shown in Wagner et al. (1984) that it is impossible, in the general case, to establish that a model is better than another. Thus, the

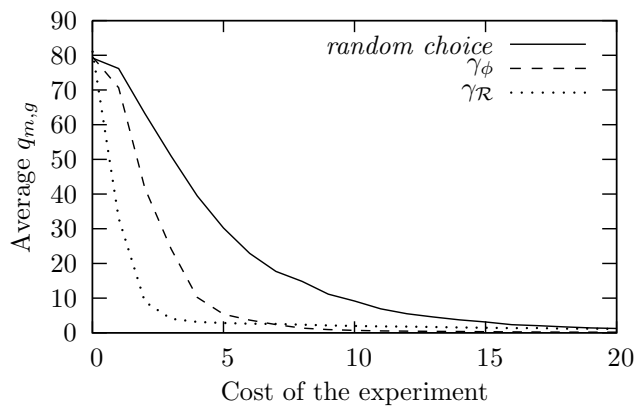


Figure 2: Quality of the heuristics is evaluated for the goal el and compared to a random choice of hypotheses.

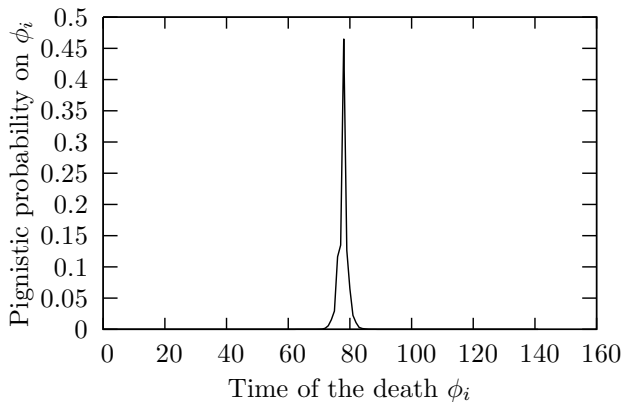


Figure 3: Pignistic probability for each hypothesis ϕ_i of Φ

multiagent model can be parametrised to use any development model. An experiment has been conducted on a real case using six different development models. Estimated result sets were merged using the Transferable Belief Model (Smets and Kennes 1994). The figure 3 shows the result of the merging after pignistic transformation. It allows to conclude that the victim died between ϕ_{55} and ϕ_{91} , and that the most probable time of death is ϕ_{78} .

This result is very interesting for at least two reasons. First, it shows that this new method allows to determine the most probable time of death more precisely than traditional methods, the agent-based model including more parameters than the classic analytic models. A discussion on the simulation results and their impact on the accuracy of PMI estimations can be found in Morvan et al. (2007). Second, a confidence interval can be formally determined from simulations results whereas in traditional methods the confidence interval estimation relies on the expert's intuition and is often questionable. As the concept of proof is crucial in criminal investigations, such an improvement is particularly interesting.

CONCLUSION

In this paper, a model of abduction is introduced and heuristics are defined to implement efficiently this model. This model is then applied in the context of Forensic Entomology. However, hypothesis sets are unidimensional in this case and large experiments should be carried on multidimensional sets. But results are encouraging and the model is being applied to develop a decision support system for cervical cancer prevention. Moreover, such a model could be used to perform intelligent exploration of model predictions or solve complex inverse problems. More generally, this study attempts to show that simulation-based reasoning models can be very useful to produce inferences in the framework of complex systems that cannot be described in logic-based languages. It seems particularly true for abductive and inductive inferences that need a higher level description than deductive (predictive) inference.

REFERENCES

- Aliseda-Llera A., 1998. *Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence*. Ph.D. thesis, Stanford University, Department of Computer Science.
- Lipton P., 2004. *Inference to the Best Explanation*. International Library Of Philosophy, Routledge, London, 2nd ed.
- McGrew T.J., 2003. *Confirmation, Heuristics, and Explanatory Reasoning*. *British Journal for the Philosophy of Science*, 54, 553–567.
- Morvan G.; Jolly D.; Dupont D.; and Kubiak P., 2007. *A Decision Support System for Forensic Entomology*. In *Proceedings of the 6th EUROSIM congress, Ljubljana, Slovenia*.
- Paul G., 1993. *Approaches to Abductive Reasoning*. *Artificial Intelligence Review*, 7, no. 2, 109–152.
- Peirce C.S., 1931. *Collected papers of Charles Sanders Peirce*, vol. 1–6. Cambridge, Harvard University Press.
- Smets P. and Kennes R., 1994. *The Transferable Belief Model*. *Artificial Intelligence*, 66, no. 2, 191–234.
- Stinner R.E.; Gutierrez A.P.; and Butler Jr G.D., 1974. *An Algorithm for Temperature-Dependant Growth Rate Simulation*. *The Canadian Entomologist*, 106, 519–524.
- Wagner T.L.; Wu H.I.; Sharpe P.J.; Schoolfield R.M.; and Coulson R.N., 1984. *Modeling insect Development Rates: A Literature Review and Application of a Biophysical Model*. *Annals of the Entomological Society of America*, 77, no. 2, 208–225.