Assessing Debris Flow Hazard by Credal Nets[*]

A. Antonucci[1], A. Salvetti[2], and M. Zaffalon[1]

[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
   Galleria 2, CH-6928 Manno (Lugano), Switzerland
   {alessandro,zaffalon}@idsia.ch
[2] Istituto Scienze della Terra (IST)
   Via Trevano, CH-6952 Canobbio, Switzerland andrea.salvetti@supsi.ch

1 Introduction

Debris flows are among the most dangerous and destructive natural hazards that affect human life, buildings, and infrastructures. Starting from the '70s, significant scientific and engineering advances in the understanding of the processes have been achieved [4, 7]. Yet, human expertise is still fundamental for hazard identification, as many aspects of the whole process are poorly understood; and the acquisition of evidence about the areas under consideration can only be done vaguely in practice, making it difficult to apply models.

This paper presents a *credal network* model of debris flow hazard for the *Ticino canton*, southern Switzerland. Credal networks [5] are imprecise-probability models based on the extension of *Bayesian networks* [9] to sets of probability mass functions (see Sec. 2.2). *Imprecise probability* [11] is a very general theory of uncertainty that measures chance and uncertainty without sharp probabilities.

The model represents expert's causal knowledge by a directed graph, connecting the triggering factors for debris flows (Sec. 3.1). Probability intervals are used to quantify uncertainty on the basis of historical data, expert knowledge, and physical theories. They are also used to carefully model the vague acquisition of evidence, which is a basis to draw credible conclusions.

The model presented here aims at supporting experts in the prediction of dangerous events of debris flow. We have made preliminary experiments

in this respect by testing the model on historical cases of debris flows. The case studies highlight the good capabilities of the model: for all the areas the model produces significant probabilities of hazard. Uncertain cases present large probability intervals.

2 Background

2.1 Debris Flows

Debris flows are composed of a mixture of water and sediment. Three types of debris flow initiation are relevant: erosion of a channel bed due to intense rainfall, landslide, or destruction of a previously formed natural dams. Prerequisite conditions for most debris flows include an abundant source of unconsolidated fine-grained rock and soil debris, steep slopes, a large but intermittent source of moisture, and sparse vegetation [3].

Several hypotheses have been formulated to explain mobilization of debris flows. Takahashi [10] modelled the process as a water-saturated inertial grain flow governed by the *dispersive stress* concept of Bagnold. In this study we adopt Takahashi's theory as the most appropriate to describe the types of event observed in Switzerland. The mechanism to disperse the materials in flow depends on the properties of the materials (*grain size*, *friction angle*), channel slope, flow rate and water depth, particle concentration, etc., and, consequently, the behavior of flow is also various.

2.2 Methods

Credal Sets and Probability Intervals

We restrict the attention to random variables which assume finitely many values (also called *discrete* or *categorical* variables). Denote by $\mathcal{X}$ the possibility space for a discrete variable X, with x a generic element of $\mathcal{X}$. Denote by $P(X)$ the mass function for X and by $P(x)$ the probability of x. Let a *credal set* be a closed convex set of probability mass functions. $\mathcal{P}_X$ denotes a generic credal set for X. For any event $\mathcal{X}' \subseteq \mathcal{X}$, let $\underline{P}(\mathcal{X}')$ and $\overline{P}(\mathcal{X}')$ be the *lower and upper probability* of $\mathcal{X}'$, respectively, defined by $\underline{P}(\mathcal{X}') = \min_{P \in \mathcal{P}_X} P(\mathcal{X}')$ and $\overline{P}(\mathcal{X}') = \max_{P \in \mathcal{P}_X} P(\mathcal{X}')$. Lower and upper (conditional) expectations are defined similarly. Note that a set of mass functions, its convex hull and its set of *vertices* (also called *extreme mass functions*) produce the same lower and upper expectations and probabilities.

Conditioning with credal sets is done by element-wise application of Bayes rule. The posterior credal set is the union of all posterior mass functions. Denote by $\mathcal{P}_X^y$ the set of mass functions $P(X|Y = y)$, for generic variables X and Y. We say that two variables are *strongly independent* when every vertex in $\mathcal{P}_{(X,Y)}$ satisfies stochastic independence of X and Y.

Let $\mathbb{I}_X = \{\mathbb{I}_x : \mathbb{I}_x = [l_x, u_x], 0 \le l_x \le u_x \le 1, x \in \mathcal{X}\}$ be a set of probability intervals for X. The credal set originated by $\mathbb{I}_X$ is $\{P(X) : P(x) \in \mathbb{I}_x, x \in \mathcal{X}, \sum_{x \in \mathcal{X}} P(x) = 1\}$. $\mathbb{I}_X$ is said *reachable* or *coherent* if $u_{x'} + \sum_{x \in \mathcal{X}, x \neq x'} l_x \le 1 \le l_{x'} + \sum_{x \in \mathcal{X}, x \neq x'} u_x$, for all $x' \in \mathcal{X}$. $\mathbb{I}_X$ is coherent if and only if the related credal set is not empty and the intervals are tight, i.e. for each lower or upper bound in $\mathbb{I}_X$ there is a mass function in the credal set at which the bound is attained [2].

The Imprecise Dirichlet Model

We infer probability intervals from data by the *imprecise Dirichlet model*, a generalization of Bayesian learning from multinomial data based on soft modelling of prior ignorance. The interval estimate for value x of variable X is given by $[\#(x)/(N+s), (\#(x)+s)/(N+s)]$, where $\#(x)$ counts the number of units in the sample in which $X = x$, N is the total number of units, and s is a hyperparameter that expresses the degree of caution of inferences, usually chosen in the interval $[1, 2]$ (see Walley's work for details [12]). Note that sets of probability intervals obtained using the imprecise Dirichlet model are reachable.

Credal Networks

A credal network is a pair composed of a directed acyclic graph and a collection of conditional credal sets. Each node in the graph is identified with a random variable (we use the same symbol to denote them and we also use "node" and "variable" interchangeably). The graph codes strong dependencies by the so-called *strong Markov condition*: every variable is strongly independent of its nondescendant non-parents given its parents. A generic variable, or node of the graph, X_i holds the collection of credal sets $\mathcal{P}_{X_i}^{pa(X_i)}$, one for each possible joint state $pa(X_i)$ of its parents $Pa(X_i)$. We assume that the credal sets of the net are *separately specified* [11]: this implies that selecting a mass function from a credal set does not influence the possible choices in others.

Denote by $\mathcal{P}$ the *strong extension* of a credal network. This is the convex hull of the set of joint mass functions $P(\boldsymbol{X}) = P(X_1, \ldots, X_t)$, over the t variables of the net, that factorize according to $P(x_1, \ldots, x_t) = \prod_{i=1}^{t} P(x_i | pa(X_i))$ $\forall (x_1, \ldots, x_t) \in \times_{i=1}^{t} \mathcal{X}_i$. Here $pa(X_i)$ is the assignment to the parents of X_i consistent with $(x_1, \ldots, x_t)$; and the conditional mass functions $P(X_i | pa(X_i))$ are chosen in all the possible ways from the respective credal sets. The strong Markov condition implies that a credal network is equivalent to its strong extension. Observe that the vertices of $\mathcal{P}$ are joint mass functions. Each of them can be identified with a Bayesian network [9], which is a precise graphical model. In other words, a credal network is equivalent to a set of Bayesian networks.

Computing with Credal Networks

We focus on the task called *updating*, i.e. the computation of $\underline{P}(X|E = e)$ and $\overline{P}(X|E = e)$. Here E is the vector of variables of the network in a known state e (the *evidence*), and X is any other node. The updating is intended to update prior to posterior beliefs about X. The updating can be computed by (i) exhaustively enumerating the vertices P_k of the strong extension; and by (ii) minimizing and maximizing $P_k(X|E = e)$ over k, where $P_k(X|E = e)$ can be computed by any updating algorithm for Bayesian networks (recall that each vertex of the strong extension is a Bayesian network).

The exhaustive approach can be adopted when the vertices of the strong extension are not too many. In general, non-exhaustive approaches must be applied as the updating problem is *NP-hard* with credal nets [6] also when the graph is a polytree.[3] In the present work the type of network, jointly with the way evidence is collected, make the exhaustive approach viable in reasonable times. Note that the exhaustive algorithm needs credal sets be specified via sets of vertices.[4]

3 The Causal Model of Debris Flows

3.1 Qualitative Influences

The network in Fig. 1 expresses the causal relationships between the topographic and geological characteristics, and hydrological preconditions, already sketched in Sec. 2.1. The causal graph is the result of an analysis of literature on debris flows and expert's considerations[5] (see [1] for details). We are mostly interested in the leaf node, which is the depth of debris likely to be transported downstream during a flood event.

3.2 Quantification of Uncertainty

Quantifying uncertainty means to complement the causal graph in Fig. 1 with probabilistic information, in order to create a global probabilistic model for the variables of the graph. This is actually done locally, by specifying the probabilistic relationship between a node and its parents.

In the application under consideration, variables are continuous apart from G, P, H, and U. Credal nets currently deal only with categorical variables, so we discretized the continuous variables. The credal sets of the net were then specified as follows. Sets of probability mass functions were specified for G, P,

---

[3]A polytree is a directed graph with the characteristic that forgetting the direction of arcs, the resulting graph has no cycles.

[4]We used the software tool *lrs* (http://cgm.cs.mcgill.ca/~avis/C/lrs.html) to produce extreme mass functions from probability intervals.

[5]The second author of this paper acted as domain expert during all the work.
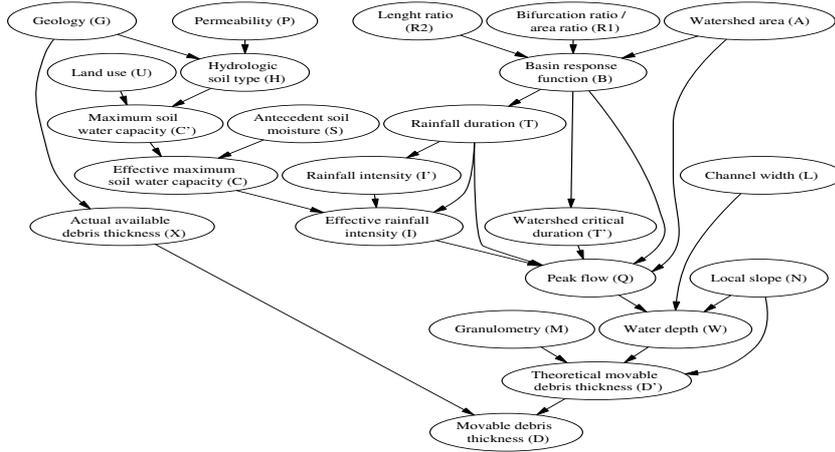
Fig. 1. The causal graph.

H, C', U, N, S, and A, by inferring probability intervals from data using the GEOSTAT database [8] and the imprecise Dirichlet model (with s=2). The expert provided interval probabilities for the nodes R_1, R_2, X, L, and M. The remaining nodes were originally continuous, depending in a deterministic way from their parents according to physical models in the literature. The discretization of these nodes naturally created precise probability mass functions for them. (The width and number of discretization classes, about 6 on average, was suggested by the domain expert.) In the following we sketch how this is done in a special case. More general cases related to the graph in Fig. 1 were reduced to such a special case.

Consider a graph made by the continuous variables X_1, ..., X_n, and Y. Let X_1, ..., X_n be root nodes in the graph, holding unconditional densities p_1, ..., p_n, and Y be direct successor of X_1, ..., X_n. Assume that the relationship between Y and its parents is described by a function of the form $y = f(x_1, \ldots, x_n)$. Now let Y be discretized using the bins $\hat{y}^{(k)} = (y_k, y_{k+1}]$. Call $\hat{Y}$ the discretized variable. We have $\hat{Y} = \hat{y}^{(k)}$ if and only if $y \in \hat{y}^{(k)}$. We use a similar notation for the discretization of the parent nodes.

We are interested in obtaining the probability mass functions for the discretized variables $\hat{X}_1$, ..., $\hat{X}_n$, and $\hat{Y}$. The case of the parent nodes is easier. For a generic parent $\hat{X}_j$, the probability of $\hat{X}_j = \hat{x}_j^{(k)}$ follows simply by integrating p_j over $\hat{x}_j^{(k)}$. In the case of $\hat{Y}$, we first have to express the function f as the degenerate conditional density $p(y|x_1, \ldots, x_n) = \delta[y - f(x_1, \ldots, x_n)]$, where δ is Dirac's delta function. Then the joint density for the considered variables becomes $p(y, x_1, \ldots, x_n) = \prod_{j=1}^{n} [p_j(x_j)] \cdot \delta[y - f(x_1, \ldots, x_n)]$. Joint probabilities for the discretized variables are computed by integrating the

joint density: $P(\hat{y}^{(k)}, \hat{x}_1^{(k_1)}, \ldots, \hat{x}_n^{(k_n)}) = \prod_{j=1}^n [\int_{\hat{x}_j^{(k_j)}} \mathrm{d}x_j \cdot p_j(x_j)] \int_{\hat{y}^{(k)}} \mathrm{d}y \cdot \delta[y - f(x_1, \cdots, x_n)]$, from which $P(\hat{y}^{(k)}|\hat{x}_1^{(k_1)}, \ldots, \hat{x}_n^{(k_n)})$ follows immediately.

4 Modelling Vague Observations

The causal model of debris flow is extended here to model vague observations of *channel depth* and *grain size* (or *granulometry*) of debris material. Indeed, channel depth and granulometry are typically known only partially, and this is a serious limit to the real application of physical theories, also considered that these two variables are very important to determine the hazard.

We focus on granulometry. (The observation of channel depth is treated in analogous way.) We model the fact than the observer may not be able to distinguish different granulometries. To this extent we add a new node to the net called O_M, which becomes parent of M. O_M represents the observation of M. There are five possible granulometries, m_1 to m_5. We define the possibility space for O_M as the power set of $\mathcal{M} = \{m_1, \ldots, m_5\}$, with elements $o_{\mathcal{M}'}$, $\mathcal{M}' \subseteq \mathcal{M}$. The observation of granulometry is set to $o_{\mathcal{M}'}$ when the elements of $\mathcal{M}'$ cannot be distinguished. $P(m|o_{\mathcal{M}'})$ is defined as follows: it is set to zero for all states $m \in \mathcal{M}$ such that $m \notin \mathcal{M}'$; and for all the others it is *vacuous*, i.e. the interval $[0,1]$ (the intervals defined this way must then be made reachable). This expresses the fact that we know that $m \in \mathcal{M}'$, and nothing else.

5 Case Studies

We validate the model in a preliminary way by an empirical study involving six areas of the Ticino canton. The credal network is fed with the information about the areas (Tab. 1), the estimated rainfall intensity on them for a return period of 10 years, and the geomorphological characteristics of the watershed. The network is thus expected to predict the probability of a debris flow event with the defined frequency level. In this way, we aim at verifying whether the network would have been a valuable tool to prevent considerable events of debris flows, which actually happened in the areas under consideration.

The results of the analysis are in Tab. 2. We use the probabilities of defined debris thickness to be transported downstream as an integral indicator of the hazard level. In the discussion below we complement this information with the classification of the areas in classes of hazard (d_1, d_2, d_3) according to the model (we do this by assuming 0-1 loss). From Tab. 2 we can see that in case 1, for example, class d_3 *dominates* the others, which means that the posterior lower probability of d_3 is strictly greater than the posterior upper probability of both d_1 and d_2. This allows us to classify case 1 as d_3. This type of dominance, sometimes called *interval dominance*, is not as strong as

Table 1. Details about the case studies.

Node	Cases					
	1	2	3	4	5	6
G	Gneiss	Porphyry	Limestone	Gneiss	Gneiss	Gneiss
U	Forest	Forest	Forest	Vegetation	Forest	Bare soil
A	0.26	0.32	0.06	0.11	0.38	2.81
N	20.8	19.3	19.3	21.8	16.7	16.7
R_1	0.9	0.6	0.7	0.9	0.9	0.8
R_2	1.5	3.5	3.5	3.5	2.3	2.1
O_M	$o_{\{m_2,m_3\}}$	$o_{\{m_1,m_2,m_3\}}$	$o_{\{m_1,m_2\}}$	$o_{\{m_3,m_4\}}$	$o_{\{m_1,m_2,m_3\}}$	$o_{\{m_3,m_4,m_5\}}$
O_L	$o_{\{l_1,l_2\}}$	$o_{\{l_2,l_3\}}$	$o_{\{l_1,l_2\}}$	$o_{\{l_2,l_3,l_4\}}$	$o_{\{l_2,l_3\}}$	$o_{\{l_3,l_4\}}$

Table 2. Posterior probability intervals of the movable debris thicknesses (cm).

Thickness	Cases					
	1	2	3	4	5	6
<10 (d_1)	[0.01,0.02]	[0.08,0.30]	[0.06,0.23]	[0.19,0.53]	[0.12,0.25]	[0.00,0.01]
10–50 (d_2)	[0.01,0.03]	[0.23,0.32]	[0.19,0.41]	[0.12,0.39]	[0.15,0.20]	[0.01,0.02]
>50 (d_3)	[0.95,0.98]	[0.46,0.69]	[0.36,0.76]	[0.34,0.45]	[0.58,0.70]	[0.97,0.99]

possible: e.g., in the same case we deduce, by using intervals, that d_1 and d_2 are mutually undominated, whereas it actually holds that d_2 dominates d_1. This can be assessed by the stronger concept of dominance called *maximality* by Walley [11]. Maximality cannot generally be read off from intervals, as interval dominance implies maximality-based dominance, by the converse is not true. We neglect details on maximality here for lack of space except that computation of maximality needs direct knowledge of the strong extension. In the following we implicity refer to maximality when we talk about dominance.

The results show that cases 1 and 6 are the most extreme in terms of high debris flow hazard level. In case 6, the relatively high upstream area and the land cover justify the results. In case 1, the slope of the source area plays probably the key role. In case 2, the probabilities of d_1 and d_2 are non-negligible, although d_3 dominates them; and the states d_1 and d_2 are mutually undominated. Similar results hold for case 5. In both cases the observations of granulometry and channel depth are identical and the upstream areas are nearly the same. The higher local slope of case 2 explains the greater lower probability of d_2. Case 4 is the most difficult to interpret, since case d_3 dominates d_2 but the system suspends judgment between d_1 and d_3. This can plausibly be explained with the very small watershed area and the low complexity level of the drainage network, probably reflected in a more uncertain estimation of geomorphological parameters. In case 3, d_2 dominates d_1, but we cannot decide between d_2 and d_3. We deduce that in case 3 the model is very sensitive to the observed channel depth and granulometry, which should be therefore carefully estimated during a field survey.

In summary, the results are recognized as compatible with the expert's understanding of the debris flow events of the case studies.

6 Conclusions

We have presented a model for determining the hazard of debris flows based on credal networks. The model unifies human expertise and quantitative knowledge in a coherent framework. This overcomes a major limitation of preceding approaches, and is a basis to obtain credible predictions, as shown by the experiments. Credible predictions are also obtained thanks to the soft-modelling made available by imprecise probability through credal nets.

The model was developed for the Ticino canton, in Switzerland. Extension to other areas is possible as the model is largely independent of the specific area. This can be accomplished by re-estimating the probabilistic information inferred from data, which has local nature.

Debris flows are a serious problem, and developing formal models can greatly help us avoiding their serious consequences. The encouraging evidence provided in this paper makes credal networks to be models of debris flows worthy of further investigation.

References

1. A. Antonucci, A. Salvetti, and M. Zaffalon. Hazard assessment of debris flows by credal networks. Technical Report IDSIA-02-04, IDSIA, 2004. http://www.idsia.ch/idsiareport/IDSIA-02-04.pdf.
2. L. Campos, J. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196, 1994.
3. J. E. Costa. *Physical geomorphology of debris flows*, chapter 9, pages 268–317. Costa, J. E. and Fleisher, P. J. - Springer-Verlag, Berlin, 1984.
4. J. H. Costa and G. F. Wiekzorek. *Debris Flows/Avalanches: Process, Recognition and Mitigation*, volume 7. Geol. Soc. Am. Reviews in Engineering Geology, Boulder, CO, 1987.
5. F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.
6. J. C. Ferreira da Rocha and F. G. Cozman. Inference with separately specified sets of probabilities in credal networks. In A. Darwiche and N. Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002)*, pages 430–437. Morgan Kaufmann, 2002.
7. R. M. Iverson, M. E. Reid, and R. G. LaHusen. Debris-flow mobilization from landslides. *Annual Review of Earth and Planetary Sciences*, 25:85–138, 1997.
8. U. Kilchenmann, G. Kyburz, and S. Winter. *GEOSTAT user handbook*. Swiss Federal Statistical Office, Neuchâtel, 2001. In german.
9. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, 1988.
10. T. Takahashi. *Debris Flow.* A.A. Balkama, Rotterdam, 1991. IAHR Monograph.
11. P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.
12. P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.