Use of Bayesian Networks as a Decision Support System for the rapid loss assessment of infrastructure systems

Pierre Gehl, Francesco Cavalieri, Paolo Franchin, Caterina Negulescu, Kristel Carolina Meza Fajardo, Kristel Meza

16TH EUROPEAN CONFERENCE ON EARTHQUAKE ENGINEERING THESSALONIKI 18 - 21 JUNE 2018

USE OF BAYESIAN NETWORKS AS A DECISION SUPPORT SYSTEM FOR THE RAPID LOSS ASSESSMENT OF INFRASTRUCTURE SYSTEMS

Pierre GEHL[1], Francesco CAVALIERI[2], Paolo FRANCHIN[3], Caterina NEGULESCU[4], Kristel MEZA[5]

ABSTRACT

This paper presents an approach for the rapid seismic loss assessment of infrastructure systems, where all probabilistic variables are modeled through a Bayesian Network (BN). While BN-based approaches have been introduced as promising tools for the risk assessment of systems, they suffer from computational issues (i.e., combinatorial explosion) that prevent their application to large real-world networks that require accurate and complex performance indicators. Therefore, a hybrid BN method is introduced here, where a preliminary Monte Carlo simulation is performed in order to generate a dataset of component damage configurations, which is used to build a simplified BN structure with only a few selected components. The most critical components are selected thanks to an unbiased importance measure computed from a random forest classification.
While the proposed approach generates an approximate BN structure that cannot provide exact probability distributions of losses, the application of Bayesian inference in a retro-analysis context (i.e., updating of loss projections given field observations immediately after an earthquake) has a lot of potential as a decision-support system for emergency responders. This method is applied to a road network in France, where evidence such as recorded ground-motions or observed damages is used to update the state of the system. The approximate BN structure has the ability to include complex system performance indicators, such as the additional travel time accounting for traffic flows. A sensitivity analysis on the component selection method and on the number of selected components demonstrates the stability of the posterior distributions, even with very few selected components.

Keywords: Probability distributions; Bridges; Road network; Situational awareness; Decision support

1. INTRODUCTION

The loss assessment of infrastructure systems has emerged as an essential aspect of the risk and resilience analysis of exposed communities (Franchin and Cavalieri, 2015). Predicting the performance loss of critical infrastructure before an event is useful to plan mitigation strategies, while a rapid loss assessment in the short-term (i.e. in the crisis period immediately following the disaster) is especially helpful for emergency responders as it contributes to situational awareness (e.g. knowledge of the areas in urgent need of basic utilities, accessibility of strategic locations, etc.). To this end, conventional approaches to model and simulate infrastructure systems include a probabilistic risk framework, where a Monte Carlo simulation is performed from the generation of earthquake events to the computation of the system performance indicators. Alternatively, Bayesian Networks (BNs) have been recently used to structure the links and statistical dependencies between the uncertain variables involved in the analysis chain (e.g. earthquake magnitude and location, ground-motion field, damage

[1]Researcher, BRGM, Orléans, France, p.gehl@brgm.fr
[2]Research Associate, La Sapienza University, Rome, Italy, francesco.cavalieri@uniroma1.it
[3]Associate Professor, La Sapienza University, Rome, Italy, paolo.franchin@uniroma1.it
[4]Researcher, BRGM, Orléans, France, c.negulescu@brgm.fr
[5]Researcher, BRGM, Orléans, France, k.mezafajardo@brgm.fr

state of infrastructure components, response of the system, etc.), thanks to the convenient use of conditional probabilities through Bayes' rule (Bensi et al., 2015). BNs may be used in a predictive (forward analysis), where all sources of uncertainties are propagated in order to obtain a probabilistic distribution of the variables of interest. On the other hand, BNs also have the ability to perform a diagnostic (backward) analysis, where the prior distribution of given variables is updated from evidence collected on fixed variables (e.g. field observations or measures). The latter property is especially relevant in the context of crisis management, since ex-ante predictive loss models may be updated thanks to the resolution of a BN with incoming evidence, thus contributing to a progressive refinement of the estimated consequences of an earthquake event (Cavalieri et al., 2017; Gehl et al., 2017).

One of the main issues preventing the application of BNs in an operational context resides in the computational complexity, which generates intractable datasets when large real-world systems are considered (Bensi et al., 2013). Moreover, formulating an exact BN, which depicts all links between variables, requires the implementation of accurate rules between the components' states (i.e. damage states of individual infrastructure systems) and the performance of the whole system. This constraint limits most BN models to a connectivity analysis, while it has been shown that capacity or serviceability analyses provide a much more accurate picture of the situation (Cavalieri et al, 2014; Hong et al., 2015). Therefore an approximate BN formulation is presented in the present study, in order to allow for a quick and efficient Bayesian updating of predictive models in near-real-time. The proposed approach is based on two distinct steps, as follows:

- Generation of a learning dataset through a Monte Carlo simulation: thousands of loss scenarios, accounting for all types of uncertainties, are sampled through infrastructure modeling and simulation tools, such as the OOFIMS platform (Franchin and Cavalieri, n.d.) developed in the FP7 SYNER-G project (Pitilakis et al., 2009-2013).
- The generated data is used to build a simplified BN formulation, where only the most influent components are kept for the estimation of the system performance. These components are selected through data mining techniques (i.e. supervised learning), and the relation between them and the system performance variable is quantified by counting the Monte Carlo outcomes for each configuration and by computing the associated conditional probability.

The main merit of this approach resides in the selection of a reduced number of influent components, which alleviates the dimensionality curse of the 'components-system' problem. Moreover, the conditional probability table of the system variable is directly built by counting combinations of events in the Monte Carlo, which allows any type of components-system relations to be represented, without necessarily being constrained by strict connectivity rules.

This two-step hybrid BN method is detailed in Section 2, where the construction of the approximate BN structure is discussed. In Section 3, the proposed approach is then applied and validated on a real-world road network in the Pyrenees area (France), where bridges are vulnerable to strong motions and road segments are exposed to earthquake-triggered landslides (more than 50 vulnerable components in total). A traffic model based on an origin-destination matrix is used to compute the Drivers' Delay (i.e., additional travel time between several points of interest) as a system performance indicator (PI). Finally, in Section 4, the inference abilities of the BN are demonstrated through various hypothetical scenarios, where field observations (e.g., ground-motion records, damage observations) are used to refine the loss estimation of the whole system.

2. LOSS ASSESSMENT OF INFRASTRUCTURE SYSTEMS WIH BAYESIAN NETWORKS

This section presents the main principles behind the modeling of infrastructure systems with BNs and details the proposed approach to build an approximate BN from Monte-Carlo simulation samples.

2.1 Modeling the Damage Probability of Spatially Distributed Components

A BN takes the form of a directed acyclic graph, which comprises edges and nodes. Nodes are classified as parents or children depending on the direction of the edges. A node without any parents is referred to as a root node. Each node represents an event or variable that may take different states

(e.g., survival or failure for a node representing an infrastructure component). The probability of each state is given by a conditional probability table (CPT), representing the probabilities given the states of the parents: in the case of a root node, the CPT becomes a table of marginal probabilities (i.e., assumed probability distribution for a given input variable). An inference is then performed on the BN when one or more nodes are observed (i.e., evidence is entered by specifying a given state) and when the probabilities of the other nodes are updated. Therefore BNs are well suited for the loss assessment of infrastructure systems, due to the convenient manipulation of conditional probabilities along the analysis chain (Bensi et al., 2011).

In the present context, the proposed BN formulation starts with the quantification of the hazard and damage events, at the level of the spatially distributed infrastructure components, as shown in Figure 1 (adapted from Bensi et., 2011; Cavalieri et al., 2017). Most of the variables are continuous and must therefore be discretized beforehand, with the exception of seismogenic areas (finite number) and components' states. The considered variables, from top to bottom, are:

- **SGZ**: root node, where each state represents one of the seismogenic zones that are susceptible to generate an earthquake event near the system;
- **M**: magnitude of the earthquake event, function of the activity parameters of the seismogenic zone;
- **Epi**: location of the earthquake event within the seismogenic zone;
- $\mathbf{R_i}$: epicentral distance for each vulnerable component *i*;
- $\overline{\mathbf{S}}_\mathbf{i}$: logarithm of the median value of the seismic intensity measure (IM) of interest, as estimated by the ground-motion prediction equation (GMPE);
- **U**: standard normal variable that is common to all sites;
- $\mathbf{V_i}$: standard normal variable that is specific to each site *i*;
- $\boldsymbol{\varepsilon_i}$: intra-event variability of the ground-motion, which is specific to each site *i*, depending on the relative contribution of the *U* and V_i Dunnett-Sobel variables (Dunnett and Sobel, 1955) that account for the spatial correlation of the ground-motion field;
- $\boldsymbol{\eta}$: inter-event variability of the ground-motion, which is common to all sites;
- $\mathbf{S_i}$: logarithmic IM at site *i*;
- $\mathbf{C_i}$: component node, with states representing the damage states of the generic component, using fragility curves to build the CPT.

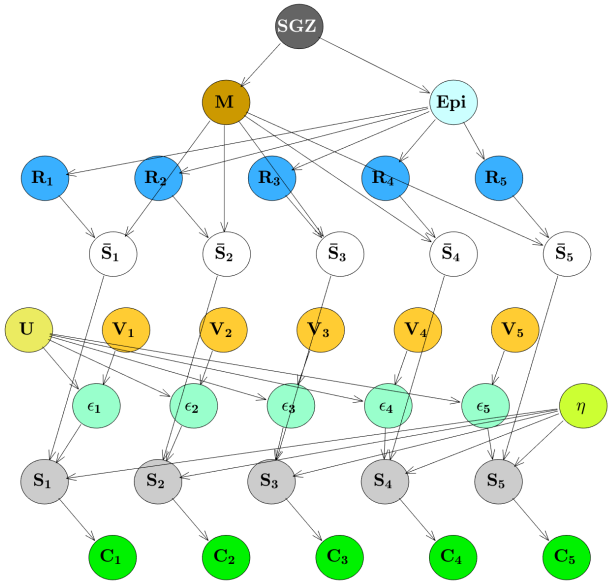

Figure 1. BN model of distributed seismic hazard, applied to a five-component example system.

The CPTs of the variables are quantified by considering established analytical and empirical models, such as GMPEs, fragility curves and earthquake recurrence laws. More details on the construction of this part of the BN are provided in Cavalieri et al. (2017). If the aim of the analysis is to estimate the

distribution of a given system performance indicator, then the BN in Figure 1 must be augmented with another set of child nodes, **SYS**, which represent the system's state given the damage state of the components nodes C_i.

2.2 Approximate Estimation of the System Performance Indicator

Directly expressing the conditional probability of the system's state as a function of all components would lead to a combinatorial explosion (i.e., the CPT has a growth rate in $O(x^n)$, n being the number of the components): such a components-system converging structure is referred to as a naïve BN formulation. Previous alternative formulations have been introduced in order to solve this issue, such as the ones based on the identification of minimum link-sets or cut-sets to decompose the problem in smaller chains of components (Bensi et al., 2013). However, it has been shown that such alternatives tend to move the computational bottleneck to other steps of the analysis, thus only slightly increasing the number of components that can be considered (Cavalieri et al., 2017).
Therefore, it is proposed here to build a reduced and approximate BN structure that accounts only for a few components to predict the system's performance. To this end, a two-step BN learning procedure is introduced, as detailed below:

1. Generation of a set of N simulated samples, through a plain Monte Carlo simulation of all the variables involved, from *SGZ* to *SYS*. The results are represented as a state matrix of size [N ; $n+1$], where each row represents the outcome of a given simulation and the n first columns represent the states of the n components in the system. The last column represents the *SYS* variable.
2. Selection of k components for the construction of the CPT of *SYS*, based on their influence on the system's performance. The CPT of *SYS* is then built only from the states of the k components, instead of all n parent nodes. The conditional probability of the discretized *SYS* to be in the state *sys*, given that the components C_i are in states c_i (for $i = 1 \ldots k$), is rewritten as in Eq. (1). The joint probabilities can then be approximated by counting the number of occurrences in the state matrix, if enough samples are generated:

$$\begin{aligned} \mathrm{P}\left(SYS = sys \middle| C_1 = c_1,...,C_i = c_i,...,C_k = c_k\right) &= \frac{\mathrm{P}\left(SYS = sys, C_1 = c_1,...,C_i = c_i,...,C_k = c_k\right)}{\mathrm{P}\left(C_1 = c_1,...,C_i = c_i,...,C_k = c_k\right)} \\ &\approx \frac{\sum_{j=1}^{N} \delta_{SYS,sys}(j) \prod_{i=1}^{k} \delta_{C_i,c_i}(j)}{\sum_{j=1}^{N} \prod_{i=1}^{k} \delta_{C_i,c_i}(j)} \end{aligned} \quad (1)$$

where N is the total number of simulated samples; and $\delta_{a,b}(j)$ is the Kronecker delta for the j^{th} sample, which takes the value 1 if $a = b$, and 0 otherwise.
In the proposed approach, a converging structure is used between the selected components and the system variable, since this configuration can be directly generated from counting the state matrix. The selection of the most critical components to include in the system's prediction is performed here through a random forest classification (Breiman, 2001), which is relevant for discrete or categorical variables. It carries out a bootstrap operation on many decision trees, so that the aggregated decision tree reduces the effect of model overfitting and provides a stable classification (e.g., reduction of the impact of components that are very rarely damaged in the Monte Carlo simulation). The bootstrap sampling is carried out on two levels, namely *(i)* on the simulation outcomes (i.e., rows of the state matrix) before each classification tree is built, and *(ii)* on the components to consider (i.e., columns of the state matrix) for each decision split in the classification tree. Using the random forest, an unbiased prediction importance estimate for each component can be retrieved in order to rank the most important components.
The proposed hybrid BN approach has the benefit of using a much smaller amount of components in order to reduce the computational complexity. Moreover, it enables any type of PI to be estimated, since the relation between the components' states and the *SYS* variables is simply obtained by counting, without the need to use any connectivity or capacity rules. However, since the approximate

BN formulation is learned from the Monte Carlo simulation, it is not able to use component configurations that are not explored by such simulation, thus providing very little benefit over conventional simulation-based approaches. On the other hand, the inference abilities of such a Bayesian framework are well suited to the diagnostic analysis of an infrastructure system immediately following an earthquake: the BN may be seen as a support tool to update initial model predictions from field observations, in order to provide a posterior distribution of the variables of interest.

3. SEISMIC LOSS ASSESSMENT OF A ROAD NETWORK IN FRANCE

This section describes the French road network used as case-study and the probabilistic loss assessment of the system, carried out in order to generate data for the BN learning.

3.1 Description of the Case-Study

The case-study area is located in the Pyrenees mountain range in the South-West of France, where a portion of a road network connecting small towns and villages is modeled. Based on the results of previous seismic risk studies (e.g., SISPYR project, www.sispyr.eu), ground shaking has the potential to affect engineering works such as bridges or even to trigger landslides on the unstable slopes that overhang some road segments. In total, the network model is composed of 219 nodes and 265 bidirectional edges: 58 edges, namely 20 bridges and 38 unstable slopes, are considered to be vulnerable to seismic hazard. For the network analysis, 10 Traffic Analysis Zones (TAZs) have been selected, corresponding either to population settlements or to entry points to the network. The road network is presented in Figure 2, together with a close-up on its central part, where most of the vulnerable components are located.

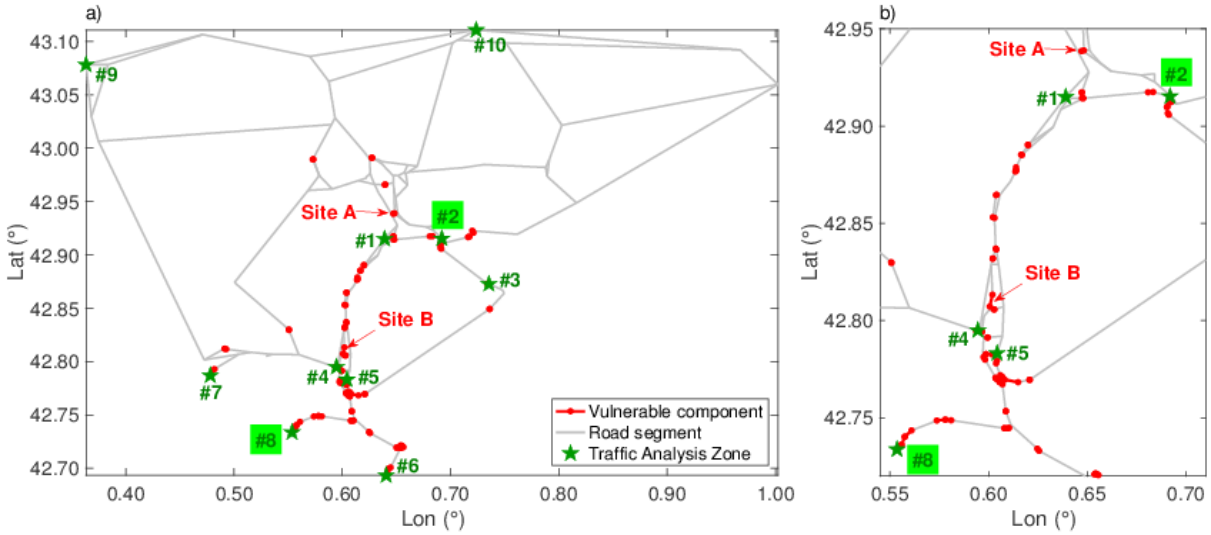

Figure 2. *(a)* Schematic view of the road network and *(b)* zoomed-in area around the sites and TAZs of interest. Sites A and B represent vulnerable components that are used as a source of field observations in the Bayesian updating. TAZs #2 and #8 respectively represent the end of the valley (ski resort) and the town of Saint-Béat, which are used to estimate trips at the local level.

The seismic hazard is modeled by following the probabilistic framework of the SYNER-G project: earthquake events are sampled from seismogenic zones surrounding the studied area, based on their activity parameters (see Table 1). The seismogenic areas have been truncated so that only the parts within 100 km to the closest vulnerable components are kept: this optimization allows more damaging earthquake events to be sampled, instead of many far-field earthquakes that would be irrelevant for the construction of the state matrix.

Table 1. Seismic activity parameters of the selected seismogenic zones (Woessner et al., 2013). λ_0 is the mean annual rate of the events in the source with magnitude M greater than the lower limit M_L, β is the magnitude slope, and M_L and M_U are the lower and upper magnitude limits of truncated Gutenberg-Richter recurrence law. The mean annual rate λ_0 has been adjusted by the ratio of the selected area (i.e., the one within 100 km of the infrastructure) on the total area of the seismogenic zone.

Zone	FRAS468	FRAS466	FRAS470	FRAS469	FRAS110	ESAS971	FRAS473
λ_0	0.0028	0.0061	0.0066	0.0053	0.0067	0.0090	0.0012
β	2.3026	2.3026	2.3026	2.3026	2.3026	2.3717	2.3026
M_L	5.5	5.5	5.5	5.5	5.5	5.5	5.5
M_U	6.8	6.8	6.8	6.8	6.5	6.8	6.8

Other modeling assumptions regarding the hazard and risk assessment are the following:

- The GMPE by Akkar and Bommer (2010) generates a spatially correlated ground-motion field at the vulnerable sites. Local site amplifications are taken into account through the specification of Eurocode 8 soil classes.
- Fragility curves are taken from the literature (Argyroudis and Kaynia, 2014), while considering a single limit state (i.e., slight/minor damage) for simplification purposes: for bridges, the median PGA is 0.12g and $\sigma_{\log PGA} = 0.44$; for unstable slopes, the median PGA is 0.16g and $\sigma_{\log PGA} = 0.40$.
- It is assumed that the occurrence of damage on a vulnerable edge corresponds to a reduction of 30% of the free-flow speed (i.e., functionality loss).
- The network performance is assessed by accounting for the traffic flow level, rather than purely in terms of connectivity or free-flow travels. To this end, an origin-destination (O-D) matrix is generated, with trips between the ten TAZs in vehicles per hour (vph). The traffic flow analysis is then carried out by reaching user equilibrium with the Frank-Wolfe algorithm.

Two system PIs are considered here, namely the global Drivers' Delay (*DD*) and a local Drivers' Delay (*LDD*), the latter considering only the travel delay between two TAZs (i.e, TAZs #2 and #8 as shown in Figure 2). *DD* is defined as the difference between the congested total travel time in damaged and normal, undamaged conditions (denoted with subscript "0"). Such total travel time is the sum of flow dependent travel times $TT(x)$ over all network edges, indexed by i, weighted by edge flows x (Shinozuka et al., 2003):

$$DD = \sum_i x_i \cdot TT_i(x_i) - \sum_i x_{0,i} \cdot TT_{0,i}(x_{0,i}) \tag{2}$$

The *LDD* PI has the same definition as *DD*, but with both summations extended over only the edges belonging to the shortest path between the two TAZs:

$$LDD\left(TAZ_{\#1}, TAZ_{\#2}\right) = \sum_{i \in \text{path}} x_i \cdot TT_i(x_i) - \sum_{i \in \text{path}} x_{0,i} \cdot TT_{0,i}(x_{0,i}) \tag{3}$$

3.2 Monte Carlo Simulation for the Selection of Components

The OOFIMS (Object-Oriented Framework for Infrastructure Modeling and Simulation) platform (Franchin and Cavalieri, n.d.) is used to model the road network and to sample 10,000 outcomes of the system's performance metrics, in terms of *DD* and *LDD*. The OOFIMS platform outputs a state matrix of size [1000 x 60], with the first 58 columns representing the components' states and the last two the performance metrics *DD* and *LDD*. This state matrix constitutes the dataset of descriptor/target variables for the creation of the random forest classification, from which unbiased performance measures are extracted in order to rank the components. As the random forest classification is specific to each system PI considered, two different sets of ten components are selected. For each PI, it is then possible to count the occurrences of the various combinations of the selected components' states, and to evaluate the probabilities of the system PI to be in a given state for each combination (see example in Table 2).

Table 2. Occurrences and probability estimation of *DD* being in the 1st discrete interval, for the five most frequent combinations of ten components, over the 10,000 outcomes of the state matrix.

ID	States (1=intact, 2=damaged) of the ten selected components										Total occurrences	Occurrences of the 1st DD interval	Probability of the 1st DD interval
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10			
1	1	1	1	1	1	1	1	1	1	1	8170	8133	0.9955
2	1	1	1	1	1	1	1	1	2	1	156	148	0.9487
3	2	1	1	1	1	1	1	1	1	1	127	0	0.0000
4	1	2	1	1	1	1	1	1	1	1	125	106	0.8480
5	2	2	1	1	1	1	1	1	1	1	90	0	0.0000
…	…	…	…	…	…	…	…	…	…	…	…	…	…

4. POST-EARTHQUAKE RAPID LOSS ASSESSMENT

This section details the approximate BN structure that is generated from the Monte-Carlo simulation, in order to demonstrate its application as a rapid loss assessment tool.

4.1 Construction of the Approximate Structure of the Bayesian Network

Once the CPTs for both system PIs have been estimated from the state matrix, the BN is built by using an exact formulation up to the component nodes (i.e., as in Figure 1), and an approximate formulation from the component nodes to the *SYS* nodes (i.e., both PIs). The resulting BN is displayed in Figure 3, where it can be seen that only ten edges converge to each *SYS* node: it comprises 355 nodes and 544 edges. In order to perform inference on this BN, the Bayes Net toolbox (Murphy, 2001) has been used, which mainly requires the CPTs and the topology of the BN to be specified. All continuous variables must be discretized beforehand, so that exact inference engines such as the junction-tree algorithm may be used. It should be noted that the size of CPTs and cliques in the junction-tree algorithm is directly linked to the number of states in the BN nodes, thus limiting the number of discrete intervals: here, with all continuous variables discretized in 10 or 20 intervals, the largest clique size generated by the junction-tree algorithm reaches a little more than 430,000,000 elements.

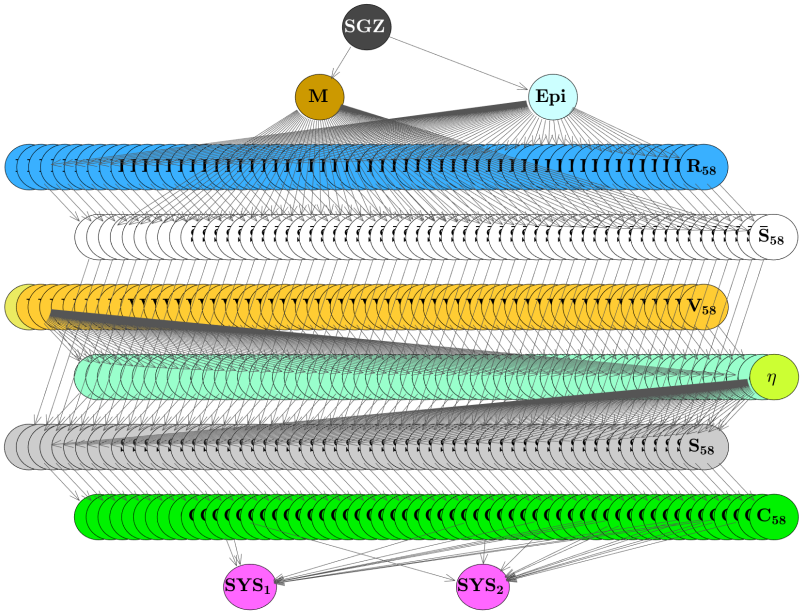

Figure 3. Layout of the t-Naïve BN formulation for the case-study, with ten components selected for each PI.

4.2 Updating of Loss Predictions through Bayesian Inference

Several inference operations have been performed on the BN in order to demonstrate its ability to account for various types of field observations and update the probability distributions of other variables. If this BN framework is to be used in the context of crisis management, the following evidences may be entered in the BN in order to update target variables (i.e., marginalized nodes) such as system PIs:

- Estimation of the earthquake magnitude and epicenter location, which is usually known within several minutes after the event;
- Measure of the ground-motion intensity at some locations by recording stations;
- Observation of damaged physical components through ground or airborne reconnaissance.

Other evidences could include the observation of some local PIs, on the condition that these loss metrics are actually measurable or observable (e.g., disruption of water flow at a given location of a water supply system). As such measurable PIs are practically unavailable in the case of road networks, only the observations at the level of the components are considered here. The proposed inference scenarios on the BN are described in Table 3, while the resulting prior and posterior distribution of all scenarios are detailed in Figure 4.

Table 3. Proposed inference scenarios for the demonstration of the BN applied to the road network. Sites A and B are shown in Figure 2, while the epicenter coordinates are [42.723°N;1.172°E].

Scenario ID	Evidence	Marginalized node
#0 (prior)	None	*LDD*, *DD*, IM_B, C_B
#1	Epicenter (R_{avg} = 49 km), M_w 6.5	*LDD*, *DD*
#2	Epicenter (R_{avg} = 49 km), M_w 6.5, C_A and C_B damaged	*LDD*, *DD*
#3	Epicenter (R_{avg} = 49 km), M_w 6.5, IM_A and IM_B high	*LDD*, *DD*
#4	Epicenter (R_{avg} = 49 km), M_w 6.5, IM_A high	IM_B, C_B

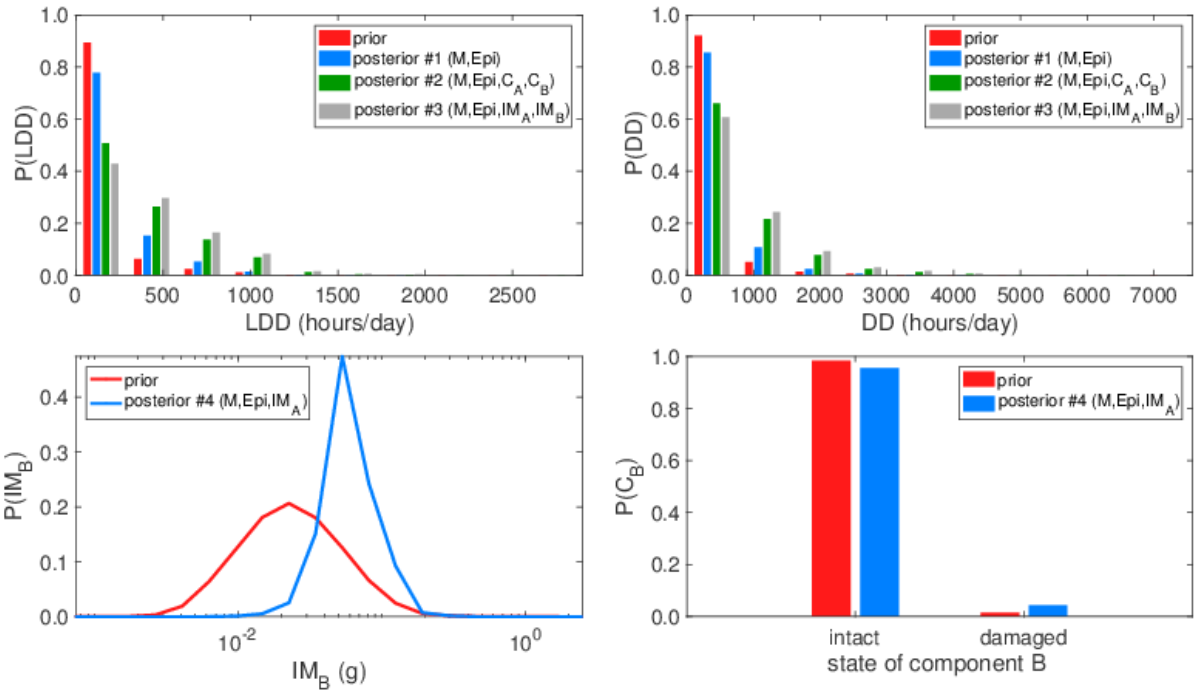

Figure 4. Prior and posterior distributions for the considered inference scenarios.

As shown in Figure 4, introducing evidence of a severe event (e.g., large magnitude, high local

intensity measures, observation of damaged components, etc.) leads to a shift of the loss distributions towards the right. It should be noted that, even though components A and B are not included in the ten components selected for the estimation of PIs *DD* and *LDD*, evidence on their damage states or the hazard intensity at their locations has a significant impact on the performance of the road network. This observation demonstrates the ability of the proposed approximate BN formulation to provide accurate estimates of the system behavior while including a reduced number of components. The observed effect is made possible by the statistical dependency between the IM_i variables (i.e., spatially correlated field), which propagates the evidence to neighboring components and finally to the system PIs (e.g., see the two bottom plots in Figure 4).

Other noteworthy observations are the following:

- The *LDD* distribution is more heavily affected by the additional evidence on C_A and C_B (i.e., difference between inference scenarios #1 and #2) than the *DD* distribution. Since *LDD* is a local PI measuring the accessibility between two TAZs, it usually involves a reduced set of very influent components, so that selecting ten components out of the total 58 provides an accurate estimation of the local performance of the network. On the other hand, *DD* is based on all inter-TAZ trips and the ten selected components are slightly less efficient to fully describe the global behavior of the network.
- The two bottom plots in Figure 4 are the result of an exact BN inference with an accurate modeling of the variables, since all the nodes involved correspond to the part of the BN where an exact formulation is used (see Figure 1). The only potential source of error lies in the discretization of continuous variables such as R_i or IM_i, which may lead to imprecise representations of the probability density functions. This is another benefit of the hybrid BN approach, where only the components-system relationship is approximate.
- Inference scenario #4 appears to have a significant impact on the distribution of the hazard intensity at site B (IM_B), yet it does not lead to a huge change in the damage distribution of C_B. However, even if the updating of the damage probability of C_B is marginal, the integration of all components at the system level provides a lever effect, where the joint damage probabilities of several components have a high impact on the network performance.

4.3 Sensitivity of the Results with respect to Component Selection

In order to investigate the stability of the selection method, the random forest classification (Breiman, 2001), is compared to other ranking algorithms, such as the Pearson correlation coefficient (as initially proposed by Cavalieri et al., 2017) and regression or classification trees (see Figure 5).

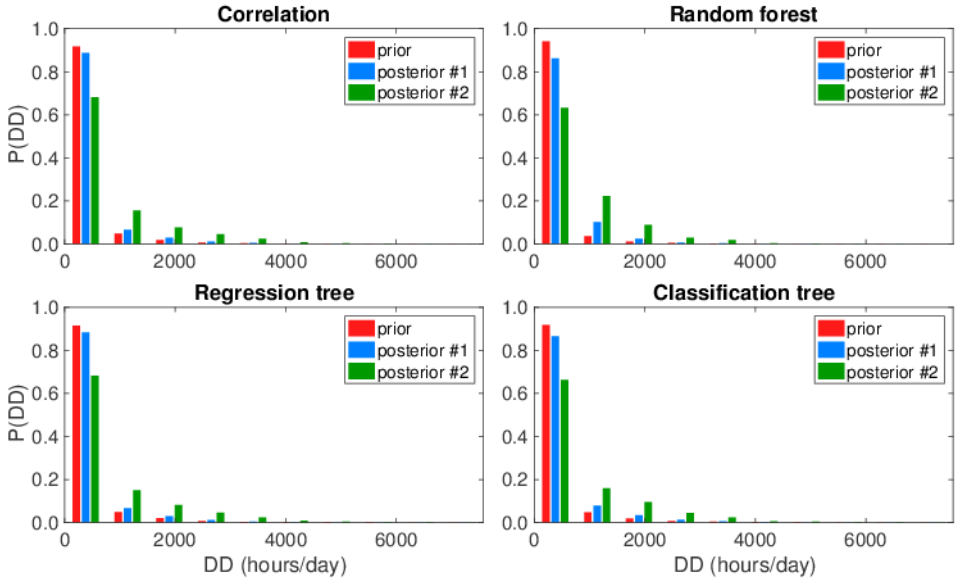

Figure 5. Sensitivity of the prior and two posterior distributions of *DD* to the component selection method.

The comparison in Figure 5 corresponds to the prior and two posterior distributions of *DD* with inference scenarios #1 and #2. The four subplots appear very similar, and this gives an estimate of the robustness of the four methods. Given the type of evidence, large gaps in probability at the first *DD* state are expected between the prior and posterior #1, as well as between the two posteriors. It is possible to note from Figure 5 that the random forest algorithm presents the largest gaps and thus allows the impact of the evidence on the PI distribution to be better captured.
The random forest algorithm for component selection involves a stochastic process and hence a variability of the solution, in the form of epistemic uncertainty. To investigate the sensitivity of the results to the ten component sequence that is output from the random forest algorithm, a total of 50 sequences have been generated and for each sequence the inferences have been performed, in terms of the prior and two posterior distributions (#1 and #2 in Table 3) of *DD*, in particular the first *DD* state; then the statistics (mean and standard deviation) of the inference results have been computed. Table 4 highlights that the highest value of standard deviation is still very low, thus confirming the robustness of the component selection via random forest.

Table 4. Sensitivity of the inference results to uncertainty in the sequence of ten components generated via random forest algorithm. The results are referred to the probability of the first state of *DD*, according to the prior distribution and two posterior distributions.

Distribution	**Inference type**		
	P[*DD*(1)]$_{prior}$	**P[*DD*(1)]$_{posterior\#1}$**	**P[*DD*(1)]$_{posterior\#2}$**
Mean	0.9234	0.7557	0.5569
St. Dev.	0.0020	0.0052	0.0128

Finally, in order to investigate the sensitivity of the solution with respect to the number of selected components, the prior and two posterior distributions (#1 and #2 in Table 4) of *DD* have been computed with an increasing number of components (Figure 6, left). Taking the values with thirteen components as "exact", it is possible to conclude that the performance is quite well captured with as low as four components. This is also evidenced by Figure 6 (right), where the normalized importance measure reaches much larger values for the first four components and presents a large decrease after the fourth one. Based on these results, the number of components for the inferences of Figure 4 has been set to ten, which is a good compromise between accuracy in the results and computational effort.

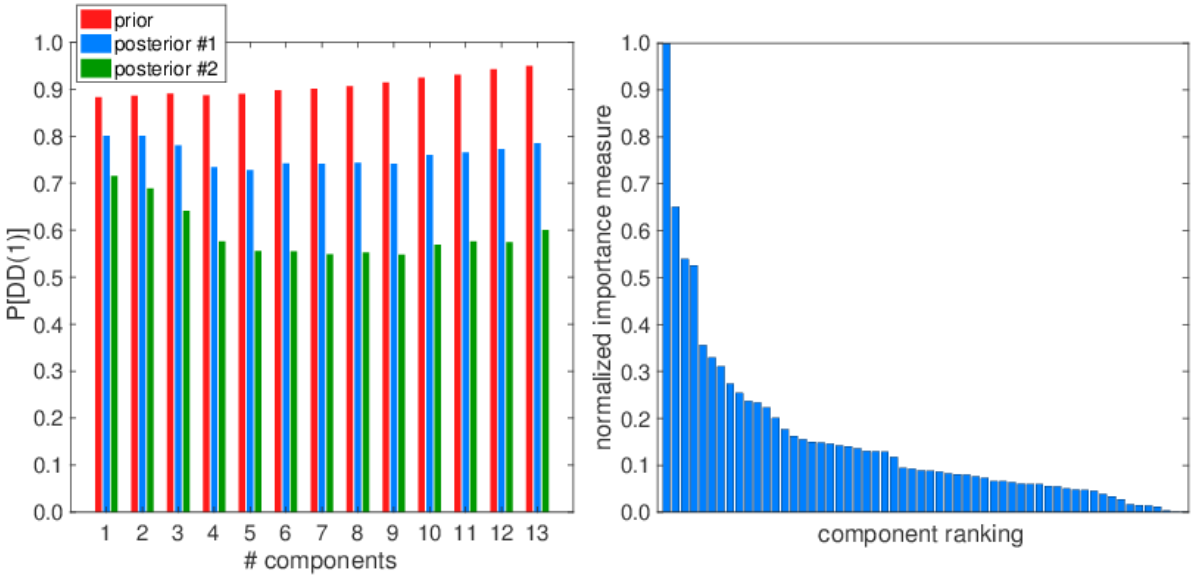

Figure 6. Sensitivity of the prior and two posterior distributions of *DD* to the number of selected components (left), and normalized importance measure of the 58 components ranked by decreasing importance (right).

The above results regarding the sensitivity to various selection parameters are specific to the present case-study, as they depend on many factors such as the network topology, the hazard distribution or the vulnerability of the components: in the present example, Figure 6 clearly indicates that the remaining components, from the fifth one on, do not play an critical role in the estimation of the quantities of interest and could be overlooked for a rough estimation of the distribution of *DD*. Therefore a similar sensitivity study should be carried out for any new case-study considered, in order to assess beforehand how many components need to be included to obtain a stable estimate. It remains feasible to perform the proposed sensitivity analyses after the Monte Carlo simulation phase (i.e., "off-line" computations ahead of any potential earthquake), in order to build a robust BN that can then be used in an operational capacity.

5. CONCLUSIONS

This paper has presented the potential benefits of using BNs for the seismic loss assessment of infrastructure systems, in complement to simulation-based approaches. The hybrid BN method relies on a preliminary Monte Carlo simulation in order to learn an approximate BN structure, characterized by a reduced number of components involved in the prediction of the system's performance. It has been shown that this approach has the potential to avoid some of the computational challenges usually associated to BNs, while being able to account for any type of performance indicators (even the *flow-based* ones).

The application of the hybrid BN approach to a road network in France has led to stable estimates of the posterior distribution of the drivers' delay measure. For this specific example, the Bayesian updating of loss probabilities based on field observations has provided satisfying results, even when considering a fraction of the vulnerable components. This encouraging observation may be explained by two main factors:

- Component failures are statistically dependent through the spatial correlation of the ground-motion field, which enables some component events to be considered as proxies for the others;
- The BN has an exact structure for all variables up to the component states, while only the system PI variable is approximately characterized in terms of linked component nodes and CPT.

The applicability of this approach to any type of infrastructure systems, however large and complex, remains to be investigated, although case-specific sensitivity studies performed on the number of selected components or the selection algorithms constitute useful tools to estimate the level of uncertainty that should be expected when studying a given area. Finally, it should be kept in mind that the use of a BN with discrete variables may also be a source of imprecision due to the discretization of continuous variables: this issue should be the subject of future investigations, with the need to develop adaptive discretization schemes that refine the probability distributions at the points of interest and optimize the number of discrete states required.

6. ACKNOWLEDGMENTS

This research has been partially supported by the internal research program PSO VULNERABILITE at BRGM, and by the Italian Civil Protection Department (DPC) through the research program Reluis-DPC 2016 task RS6.

7. REFERENCES

Akkar, S, Bommer, JJ (2010). Empirical equations for the prediction of PGA, PGV and spectral accelerations in Europe, the Mediterranean region and the Middle East. *Seismological Research Letters*, 81(2):195-206.

Argyroudis S, Kaynia AM (2014). Fragility functions of highway and railway infrastructure. In: Pitilakis K, Crowley H, Kaynia AM (eds) SYNER-G: Typology definition and fragility functions for physical elements at seismic risk. *Geotechnical, Geological and Earthquake Engineering* 27, Springer Netherlands, pp 299-326.

Bensi M, Der Kiureghian A, Straub D (2015). Framework for post-earthquake risk assessment and decision making for infrastructure systems. *ASCE-ASME Journal of Risk Uncertainty and Engineering Systems, Part A: Civil Engineering*, 1:1-17.

Bensi M, Der Kiureghian A, Straub D (2013). Efficient Bayesian network modeling of systems. *Reliability Engineering and System Safety*, 112:200-213.

Bensi M, Der Kiureghian A, Straub D (2011). A Bayesian Network methodology for infrastructure seismic risk assessment and decision support. *Technical Report 2011/02*, Pacific Earthquake Engineering Research Center, Berkeley, California.

Breiman L (2001). Random Forests. *Machine Learning*, 45:5-32.

Cavalieri F, Franchin P, Gehl P, D'Ayala D (2017). Bayesian networks and infrastructure systems: Computational and methodological challenges. In: Gardoni P (ed) Risk and reliability analysis: Theory and applications. *Springer*, pp. 385-415.

Cavalieri F, Franchin P, Buriticá Cortés JA, Tesfamariam S (2014). Models for seismic vulnerability analysis of power networks: comparative assessment. *Computer-Aided Civil and Infrastructure Engineering*, 29:590-607.

Dunnett CW, Sobel M (1955). Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t-distribution. *Biometrika*, 42:258-260.

Franchin P, Cavalieri F (2015). Probabilistic assessment of civil infrastructure resilience to earthquakes. *Computer-Aided Civil and Infrastructure Engineering*, 30:583-600.

Franchin P, Cavalieri F. n.d. OOFIMS, Object-Oriented Framework for Infrastructure Modeling and Simulation. Available from: sites.google.com/a/uniroma1.it/oofims/ [accessed 27 February 2018].

Gehl P, Cavalieri F, Franchin P, Negulescu C (2017). Robustness of a hybrid simulation-based/Bayesian approach for the risk assessment of a real-world road network. *Proceedings of the 12th International Conference on Structural Safety and Reliability*, 6-10 August, Vienna, Austria.

Hong L, Ouyang M, Peeta S, He X, Yan Y (2015). Vulnerability assessment and mitigation for the Chinese railway system under floods. *Reliability Engineering and System Safety*, 137:58-68.

Murphy K (2001). The Bayes Net toolbox for Matlab. *Computer Science and Statistics*, 33:1024-1034.

Pitilakis K and the SYNER-G consortium (2009-2013). SYNER-G: Systemic Seismic Vulnerability and Risk Analysis for Buildings, Lifeline Networks and Infrastructures Safety Gain. Available from: www.vce.at/SYNER-G/ [accessed 27 February 2018].

Shinozuka M, Murachi Y, Dong X, Zhou Y, Orlikowski MJ (2003). Seismic performance of highway transportation networks. *Proceedings of the China-US Workshop on Protection of Urban Infrastructure and Public Buildings against Earthquakes and Manmade Disasters*, 21-22 February, Beijing, China.