Approximate Bayesian Network Formulation for the Rapid Loss Assessment of Real-World Infrastructure Systems

Pierre Gehl , Francesco Cavalieri , Paolo Franchin

PII: S0951-8320(17)31337-6
DOI: 10.1016/j.ress.2018.04.022
Reference: RESS 6141

To appear in: *Reliability Engineering and System Safety*

Received date: 14 November 2017
Revised date: 23 April 2018
Accepted date: 27 April 2018

Please cite this article as: Pierre Gehl , Francesco Cavalieri , Paolo Franchin , Approximate Bayesian Network Formulation for the Rapid Loss Assessment of Real-World Infrastructure Systems, *Reliability Engineering and System Safety* (2018), doi: 10.1016/j.ress.2018.04.022

Highlights

- The proposed Bayesian Network can treat large systems with complex performance metrics
- A random forest algorithm is adopted for a stable selection of important components
- The influence of evidenced components is enhanced by a recursive building algorithm
- A similarity measure ensures the robustness of the off-line Monte Carlo simulation
- The method is applied to a real-world road network, with a sensitivity analysis

Approximate Bayesian Network Formulation for the Rapid Loss Assessment of Real-World Infrastructure Systems

Pierre Gehl*[1], Francesco Cavalieri[2], Paolo Franchin[2]

[1] Risks and Prevention Division, BRGM, Orléans, France

[2] Department of Structural and Geotechnical Engineering, Sapienza University of Rome, Rome, Italy

* = corresponding author: p.gehl@brgm.fr

Abstract

This paper proposes to learn an approximate Bayesian Network (BN) model from Monte-Carlo simulations of an infrastructure system exposed to seismic hazard. Exploiting preliminary physical simulations has the twofold benefit of building a drastically simplified BN and of predicting complex system performance metrics. While the approximate BN cannot yield exact probabilities for predictive analyses, its use in backward analyses based on evidenced variables yields promising results as a decision support tool for post-earthquake rapid response. Only a reduced set of infrastructure components, whose importance is ranked through a random forest algorithm, is selected to predict the performance of the system. Further, owing to the higher importance of evidenced nodes, the ranking method is enhanced with a recursive evidence-driven BN-building algorithm, which iteratively inserts evidenced components into the subset identified by the random forest algorithm. This approach is applied to a French road network, where only 5 to 10 components out of 58 are kept to estimate the distribution of system performance metrics that are based on traffic flow. Sensitivity studies on the number of selected components, the number of off-line simulation runs and the discretization of variables reveal that the reduced BN applied to this specific example generates trustworthy estimates.

Keywords: Bayesian networks; seismic risk; decision support; road network; Bayesian learning; system performance

1. Introduction

The performance of critical infrastructure plays an essential part in the disaster management phase following an earthquake, as demonstrated by recent events such as the 2016 Kaikoura (New Zealand) earthquake (M_w 7.8), which cut off many settlements. The development of rapid response systems, able to update damage predictions in near-real time from field observations, represents one of the main challenges of disaster reduction efforts [1]. Thanks to their inference abilities, Bayesian Networks (BNs) offer appropriate mathematical tools for the rapid updating of projected loss distributions, as an input to decision support systems [2].

BNs model the dependencies between variables through directed edges and conditional probability tables (CPTs, in the case of a BN with discrete variables), which provide conditional probabilities given the states of parent variables. However, in the specific case of infrastructure systems, physical components are usually interconnected and most of them contribute to the performance of the system. As a result, evaluation of a system performance measure (S) based on the combination of the states of *n* components is one typical case of dimensionality curse, as the CPT size grows exponentially with the number of components ($O(c^n)$ with $c \geq 2$). Bensi et al. [3,4] have explored various strategies based on the BN's topology in order to alleviate this issue: they advocate the grouping of components into parallel or series sub-systems through the identification of minimum link sets or cut sets, thus limiting the amounts of edges converging towards a given node. However, while the aforementioned BN formulations enable the number of components to be increased to some extent, Cavalieri et al. [5] have shown that memory issues start to appear as soon as a couple of dozen binary components are considered, even when efficiently coalescing the intermediate nodes that represent survival or failure sequences. Moreover, these BN formulations require the preliminary identification of all minimum link sets and cut sets, usually through recursive algorithms that examine all possible connectivity paths in the considered network topology. Such a task quickly becomes overwhelming in terms of computation time when studying large systems.

Alternative BN frameworks have also been proposed, such as the use of a compression algorithm in order to reduce the memory storage space of the CPT of the system performance [6,7]. Those methods rely on a simple converging structure from the components to the system node (i.e., naïve formulation), since the CPT compression help managing a large number of parent nodes. The variable elimination inference algorithm, coupled with a careful ordering of the component nodes, is used in order to avoid repeating compression and recompression operations and to reduce the computation time. However, most of the considered cases involve independent components (i.e., component nodes are root nodes in the BN), while the case of dependent components is just addressed with a single root node governing the statistical correlation between components. Therefore, the application of the compression algorithm to a real-world system that is exposed to a spatially distributed ground-motion field, where the seismic intensity affecting the state of each component is linked to the intensity values at all other components' locations through a network of several intermediate variables, would require more developments and a cumbersome optimization of the variable elimination algorithm. On the other hand, Pozzi and Der Kiureghian [8] have investigated Gaussian BNs, which consist of continuous variables and have therefore the merit of greatly reducing CPT sizes. However, this approach requires all variables to be represented by a Gaussian distribution, which is not the case of the components' states. This strong limitation prevents the use of exact inference algorithms, and approximate inference engines such as importance sampling or Gibbs sampling have to be used instead. In order to use exact inference to treat the case of discrete children of continuous parents, it is possible to discretize all the continuous variables, as done by Hosseini and Barker [9] in modeling the resilience capacity of an inland port. Their BN is composed by Boolean and continuous variables, the latter being modeled with a truncated normal distribution that is then discretized.

Most of the aforementioned BN approaches are focused on connectivity-based performance measures only, since such a framework allows for straightforward rules to be extracted (e.g., parallel or series assembly of components). Still, recent studies [10,11] have highlighted the major influence of the measure type on the accuracy of the estimation of a system's performance. It has been shown that flow-based measures offer a much more precise picture of the state of the system than metrics based

on connectivity only. The work by Cavalieri et al. [5] has constituted a first attempt to overcome this issue, by adopting a two-step BN learning procedure. First, Monte Carlo (MC) simulations are performed in order to generate samples of the variables' states, in the form of a state matrix. Then, the most influential components, with respect to a given system performance of interest, are selected in order to build an approximate BN formulation that only includes a reduced number of nodes in the components-system converging structure. This method enables any type of measure to be considered, including flow-based ones, since the relation between the components and the system variables is directly obtained from the simulation results. It also leads to a reduction in the complexity of the BN, so that larger and more complex systems may be treated: in other words, the approximate BN formulation models a complex system through a simplified structure, thus reducing the problem to a computationally tractable case.

Therefore, this paper builds upon the idea of approximate BN formulation introduced by Cavalieri et al. [5], with the objective of improving it and demonstrating its feasibility in the case of a real-world system. While this approximate BN method has shown promising results when applied to a virtual infrastructure system [5], several points still need to be addressed, such as the most adequate importance measure for the selection of components, the number of components required in order to obtain stable inference results, the discretization scheme for continuous variables, the number of off-line MC simulation runs that lead to a reliable BN or the appropriateness of accounting only indirectly for evidence on components not directly linked to the system performance. The successive steps in the construction of the approximate BN are provided in Section 2, where the adopted supervised BN learning algorithm is detailed. An importance measure based on a random forest analysis [12] of the Monte Carlo outputs is introduced as a more robust alternative to linear Pearson's correlation. Moreover, a recursive evidence-driven BN-building algorithm is introduced, which iteratively replaces the components connected to each system variable of interest, mixing components with evidence and components from the random forest algorithm, while keeping their total number always within computable limits. Then Section 3 describes the real-world infrastructure system, i.e. a road network connecting a few towns in the Pyrenees mountain range in France, which is used for the application of

the BN framework. Finally, the BN is applied to the road network in Section 4, where various inference scenarios are demonstrated, through the inclusion of field observations that are relevant to potential disaster management operations. Section 4 also reports several sensitivity studies, which aim to evaluate the robustness of the inference scenarios with respect to the number of selected components, the number of simulation runs and the discretization of continuous variables.

2. Strategy for the BN Modeling of Real-World Infrastructure Systems

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. This BN structure is similar to the one adopted by Bensi et al. [3] and Cavalieri et al. [5], except that a Z node has been added in order to model the possibility of sampling earthquake events from different seismogenic zones. Indeed, apart from this generalization, the BN model by Bensi et al. works very well for the upper portion related to seismic hazard and even to the level of components' damage state. It is the bottom system-level portion that has limitations in dealing with larger-size systems and flow-based performance measures. This paper deals with a better model for this bottom portion of the BN.

Most of the variables are continuous and must therefore be discretized beforehand, with the exception of the seismogenic zone node (finite number of states, or zones) and components' states. Therefore, the considered variables, from top to bottom, are:

a) Z (*discrete*): root node, where each state represents one of the seismogenic zones that are susceptible to generate an earthquake event near the system (these areas have been discretized beforehand as a result of probabilistic seismic hazard assessment – e.g. Woessner et al. [13]);
b) M (*continuous*, *discretized*): magnitude of the earthquake event, function of the activity parameters of the seismogenic zone;
c) E (*continuous*, *discretized*): location of the earthquake event within the seismogenic zone (point-source model);
d) R_i (*continuous*, *discretized*): epicentral distance for each vulnerable component *i*;

e) $\bar{I}_i$ (*continuous*, *discretized*): logarithm of the median value of the seismic intensity measure (IM) of interest, as estimated by the ground-motion prediction equation;

f) U (*continuous*, *discretized*): standard normal variable that is common to all sites (first part of the Dunnett-Sobel decomposition [14] of the intra-event error term);

g) V_i (*continuous*, *discretized*): standard normal variable that is specific to each site *i* (second part of the Dunnett-Sobel decomposition [14] of the intra-event error term);

h) ε_i (*continuous*, *discretized*): 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 [14] that account for the spatial correlation of the ground-motion field;

i) η (*continuous*, *discretized*): inter-event variability of the ground-motion, which is common to all sites;

j) I_i (*continuous*, *discretized*): logarithmic IM at site *i*;

k) C_i (*discrete*): component node, with states representing the damage states of the component, using fragility curves to build the conditional probability table.

The CPTs of the variables are quantified by considering established analytical and empirical models, such as, e.g., ground motion prediction equations (GMPEs) for $p(I|M,R,\varepsilon,\eta)$, fragility curves for $p(C|I)$ and earthquake recurrence laws for $p(M|Z)$. More details on the construction of this part of the BN are provided in Cavalieri et al. [5], which adopts the Bensi et al. [3] approach.

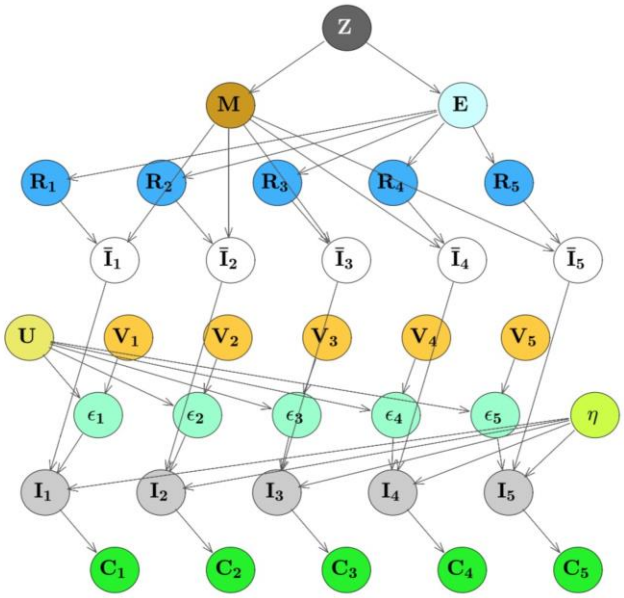

Figure 1. BN model of distributed seismic hazard, applied to a five-component example system.

As mentioned in Section 1, the lower portion of the BN, i.e. the definition of the system performance measure S as a function of the components' states C_i, is here achieved by adopting a converging structure from the components to the system node (i.e., naïve formulation). This choice is guided by the findings of Cavalieri et al. [5], who have shown that, for a reduced number of components (e.g., around a dozen), a naïve formulation remains the most computationally efficient strategy when using a junction tree algorithm. Therefore, in order to keep the number of components to a reasonable amount, it is proposed to select only the most 'critical' components, i.e. the ones that provide the most accurate conditional distribution of the S variable. To this end, a three-step BN learning procedure is introduced, as detailed below:

1. Generation of a set of N samples, through a Monte Carlo-like simulation of all the variables involved, from Z to S. The subset of results of interest for learning the BN can be represented as a state matrix of size [N ; n+m], where each row represents the outcome of a one simulation run, the first n columns represent the states (C) of the n components in the system, and the last m columns represent the m system performance measures (S) of interest. This is the most computationally intensive step and is performed only once off-line.

2. Ranking of the n components based on their influence on each system's performance measure S (m distinct rankings). This is not a computationally intensive task, and in any case it is performed only once, off-line, in order to establish the most influential components for each system measure, in the absence of any evidence (i.e. when all components' states are unobserved). The use of random forest classification as a ranking method is detailed below.
3. Selection of k components for the construction of the CPT of each of the S variables, initially based on their influence on the system's performance as established in Step 2. The CPT of S is then built only from the states of the k components, instead of all n parent nodes. The conditional probability of the discretized S to be in state s, given that the components C_i are in states c_i (for $i = 1…k$), is evaluated as in Eq. (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. The joint probabilities can then be approximated by counting the number of occurrences in the state matrix, if enough samples are generated. This step is the least computationally intensive and is repeated whenever evidence is obtained on a component initially not included among the first k in the ranking. When this happens, the component is added to the list of n_e components with evidence and the remaining k-n_e components are taken from the corresponding initial ranking from Step 2 (i.e. each time a component outside the set receives evidence, it replaces the least important component in the set). If evidence is collected on components already in the set, or on other variables such as the event magnitude, epicenter location, intensity at a site, etc. the BN stays the same.

$$\mathrm{P}\left(S=s \middle| C_1=c_1,...,C_i=c_i,...,C_k=c_k\right)=\frac{\mathrm{P}\left(S=s,C_1=c_1,...,C_i=c_i,...,C_k=c_k\right)}{\mathrm{P}\left(C_1=c_1,...,C_i=c_i,...,C_k=c_k\right)} \approx \frac{\sum_{j=1}^{n_s}\delta_{S,s}(j)\prod_{i=1}^{k}\delta_{C_i,c_i}(j)}{\sum_{j=1}^{n_t}\prod_{i=1}^{k}\delta_{C_i,c_i}(j)} \quad (1)$$

The resulting structure, referred in Cavalieri et al. [5] as the *thrifty-naïve* or *t-Naïve* formulation (Figure 2), presents the double merit of *(i)* using a much smaller amount of components in order to reduce the computational complexity, and of *(ii)* enabling any type of system performance measure S to be estimated, since the relation between the components' and system's states is simply obtained by counting, without the need to use any connectivity or capacity rules. However, since the *t-Naïve* formulation is learned from Monte Carlo simulations, it is not able to explore component configurations that are beyond the solution space discovered by the simulations. As a result, the application of the proposed method to predictive analyses (i.e., forward inference) does not provide much benefit when compared to more conventional Monte Carlo sampling. On the other hand, the inference abilities of such a Bayesian framework are well suited to the diagnostic analysis (i.e., backward inference) of an infrastructure system immediately following an earthquake: initial model predictions are updated from field observations in order to provide a posterior distribution of the variables of interest.

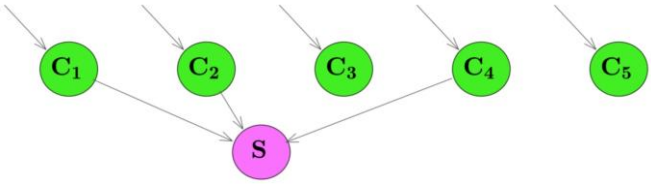

Figure 2. Example of a *t-Naïve* formulation when three components out of five are selected.

While Cavalieri et al. [5] have used the Pearson correlation between the components' states and the system states in order to rank the importance of each component, it is proposed here to use a random forest classification [12], which is more suited to discrete or categorical variables. This algorithm generates a set of single classification models such as decision trees [15]. In the case of categorical variables (i.e., damage states of the components C_i), a decision tree is built by progressively splitting the domain space $\{C_i\}_{i=1..n}$ until homogeneous regions with respect to the target variable (i.e., the state of the performance measure S) are created. The Gini index is computed at each split node of the tree in order to decide which input variable C_i has to be split next, thus creating a new set of branches (see

Figure 3). Before any splits are carried out, the Gini index of the target variable S is expressed as follows [15]:

$$GINI(0) = 1 - \sum_{j=1}^{m} \left[P\left(S = s_j\right) \right]^2 \tag{2}$$

where *m* is the number of states of S. Then, this index has to be recomputed for each candidate C_i, at the next potential split node *t* (*t* > 0):

$$\begin{aligned} GINI(C_i, t) &= GINI(t-1) - P(C_i = \bar{c}_i) \cdot GINI(t^-) - P(C_i = c_i) \cdot GINI(t^+) \\ &= GINI(t-1) - P(C_i = \bar{c}_i) \cdot \left[1 - \sum_{j=1}^{m} P(S = s_j | C_i = \bar{c}_i)^2 \right] - P(C_i = c_i) \cdot \left[1 - \sum_{j=1}^{m} P(S = s_j | C_i = c_i)^2 \right] \end{aligned} \tag{3}$$

where *GINI*(*t*-1) represents the Gini index of the target variable at the previous split node of the given branch, and *GINI*(t^-) and *GINI*(t^+) represent the Gini index given the state of component C_i at split node *t* (*failure* and *survival*, respectively).

This computation is carried out for all possible components C_i and split nodes *t*, and the combination that yields the smallest Gini index is selected to create the next split of the classification tree.

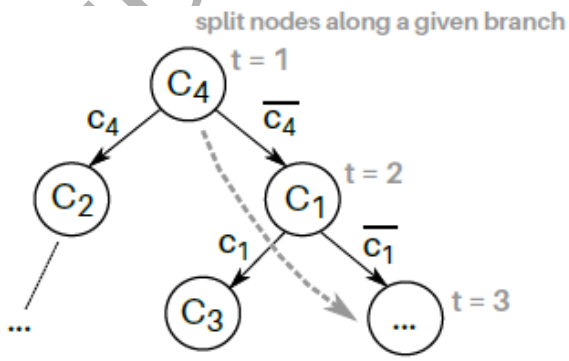

Figure 3. Construction of a classification tree, with the first split made on the states of component C_4.

The principle of the random forest classification algorithm relies on the bootstrap sampling of numerous classification trees (see Figure 4), with the aim of generating a stable classification and reducing model overfitting (e.g., reduction of the impact of components that are very rarely damaged in the Monte Carlo simulations). The bootstrap sampling is carried out on two levels, namely *(i)* on the

simulation outcomes (i.e., removal of some rows of the state matrix) before each classification tree is built, and *(ii)* on the components to consider (i.e., removal of some columns of the state matrix), for each decision split in the classification tree. For each classification tree generated, a classification error *Err* is computed, by counting the amount of misclassifications that are found when applying the classification tree to the sub-set of data that has not been included in the bootstrap sample (i.e., out-of-bag sample of size *N-N'*, if *N'* is the number of simulation outcomes selected in the bootstrap):

$$Err = \frac{1}{N - N'} \cdot \sum_{i=1}^{N-N'} \left[1 - \delta_{s_{CT}, s_{OOB}}(i) \right] \tag{4}$$

where $\delta_{a,b}(i)$ is the Kronecker delta for the i^{th} out-of-bag sample, s_{CT} is the state of S as predicted by the classification tree, and s_{OOB} is the actual state of S.

The generation of the random forest is achieved through the TreeBagger function in MATLAB [16], which creates bootstrap samples of classification trees. This algorithm is also able to provide an unbiased prediction importance estimate for each component, which may then be used to build a straightforward ranking of the most important components. This predictor importance measure corresponds to the difference between the actual error rate as computed in Eq. (4) and the error rate obtained when permuting the values taken by a given component in the state matrix [17]: a large error rate means a large variation in the target variable distribution due to the permutation, and consequently a large influence of this component on the performance measure S.

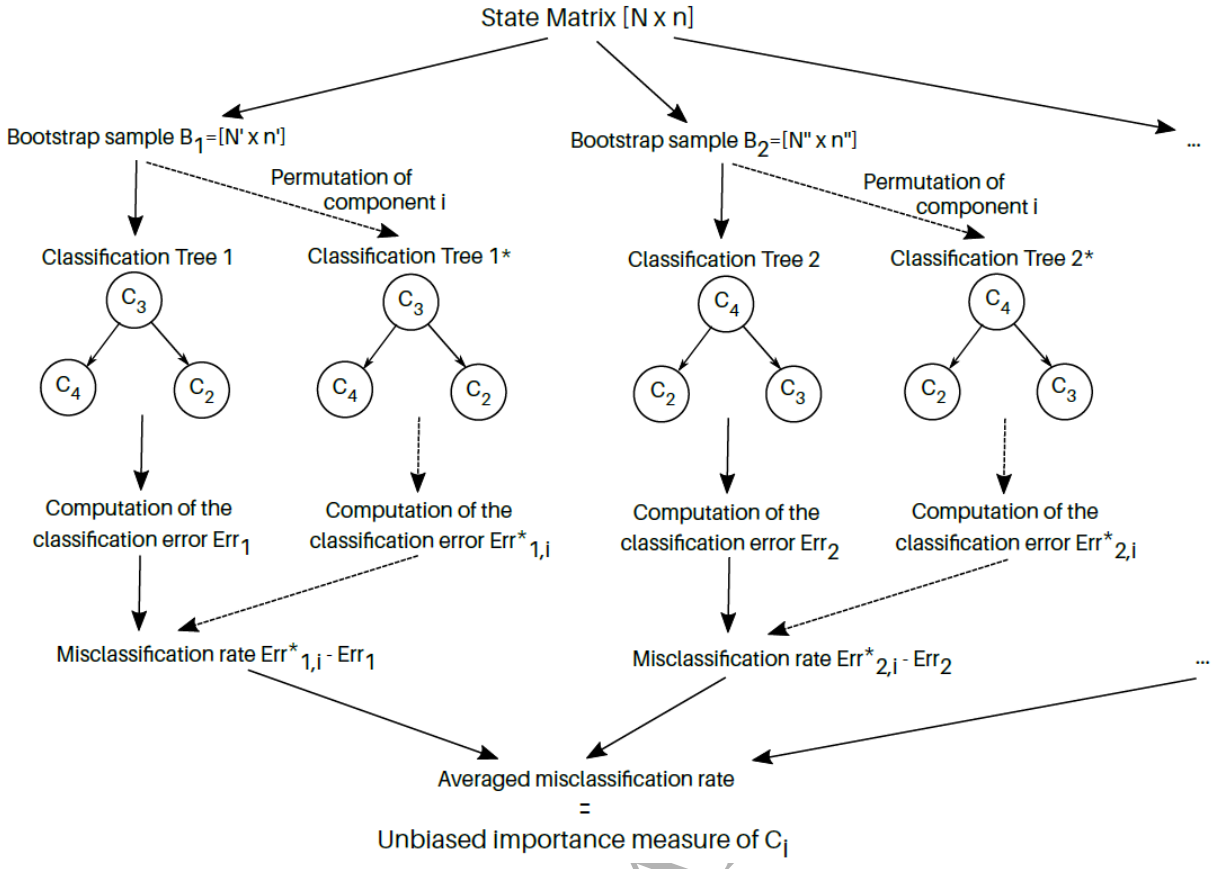

Figure 4. Construction of a random forest and computation of the unbiased importance measure, using boostrap samples of the state matrix to generate classification trees.

Alternative learning strategies have been discussed in the literature, such as the *K-2* algorithm [18], which is a form of unsupervised BN learning that creates dependency links between component nodes (i.e., no a priori assumption of the BN structure). This approach, however, is not compatible with the case of a seismic risk analysis, for which the physical relations between the hazard intensity and the states of the components must be fully explicated in the BN, in order to respect a principle of causality between physical parameters. Therefore, directly counting the state matrix in order to build the system CPT remains the most appropriate strategy in the present case, since all BN nodes correspond to engineering models [19]. Albeit out of the scope of the present study, elaborate supervised BN learning algorithms are still worth investigating, by setting for instance a prior probability of zero to all tentative BN structures that do not present a converging structure.

Finally, it is important to remark how, of the improvements with respect to [5], the introduction of a recursive evidence-driven BN-building procedure with Step 3 represents a major advancement. In the *t-Naïve* formulation, where some links are trimmed, evidence on the state of components excluded from the subset of the most influential ones, such as, e.g., components C_3 and C_5 in Figure 2, is only indirectly affecting the state of the system. The importance of the components formalized through the initial ranking in Step 2 is based on the assumption that all components are on an even starting level, affected by uncertainty. Evidence, however, is highly informative and its effect showed to be different when propagated to S indirectly, through the intensity variables (I) and then the states of other components (C), or directly from C to S. For this reason, including always the components with a known state in the subset greatly improves the performance of the approximate BN.

3. Application: Road Network in the Pyrenees (France)

This section describes the case-study that is used for the demonstration of the proposed enhanced *t-Naïve* BN formulation. It is adapted from that presented in Gehl et al. [20].

3.1. Presentation of the Case-Study

The case-study area is located in the South of France, along the border with Spain. The small towns and villages within this area are connected through a set of departmental roads that are mainly running along the steep valleys of the Pyrenees mountain range. Seismic hazard is a potentially disruptive threat, since the area is characterized by an average seismicity level according to the French seismic zonation. The region has been indeed the object of previous seismic risk studies (e.g. SISPYR, www.sispyr.eu, or ISARD projects). 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 road 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 5, together with a close-up on its central part, where most of the vulnerable components are located.

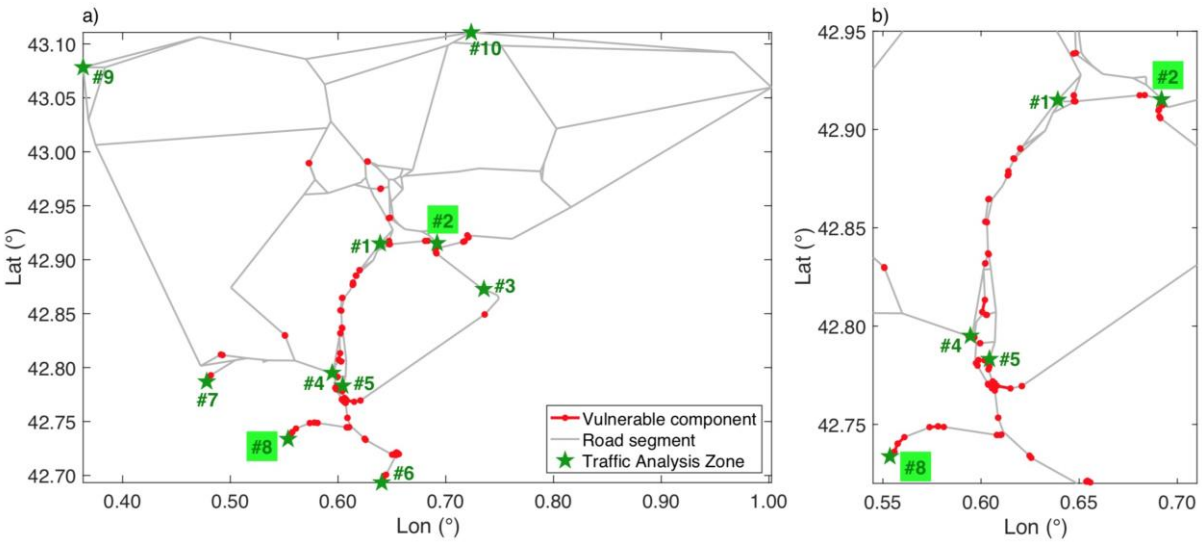

Figure 5. a) Schematic view of the road network and b) zoomed-in area around the sites and TAZs of interest. TAZs #2 and #8 respectively represent the end of the valley (ski resort) and the town of Saint-Béat, which are used to estimate trips at the local level.

3.2. Modelling Assumptions for the Supporting Monte Carlo Simulation

In order to model the seismic hazard in the area and to cover the spatial extent of the vulnerable components, seven seismogenic zones have been selected, whose activity is characterized in terms of the parameters of the truncated Gutenberg–Richter recurrence law: λ_0 (i.e., the mean annual rate of the events in the source with magnitude greater than the lower limit M_L), magnitude slope β, lower and upper magnitude limits M_L and M_U. The parameter values in Table 1 have been retrieved from Woessner et al. [13]. The seismogenic areas have been truncated so that only the parts within 100 km to the closest vulnerable components are kept (see the general layout of the areas in Figure 8): this optimization allows for more damaging earthquake events to be sampled, instead of many far-field earthquakes that would be irrelevant for the construction of the state matrix. 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.

Table 1. Seismic activity parameters of the selected seismogenic zones.

Zone #	1	2	3	4	5	6	7

λ_0	0.0028	0.0061	0.0066	0.0053	0.0067	0.0090	0.0012
β	2.303	2.303	2.303	2.303	2.303	2.372	2.303
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

A spatially correlated ground motion field is generated over the area of interest by using the GMPE by Akkar and Bommer [21], as extended by Bommer et al. [22] to periods below 0.05 s. Local site amplifications are also accounted for by the GMPE, through the specification of Eurocode 8 soil classes for the vulnerable sites. Given the illustrative character of the application, only one limit state is considered for the vulnerable edges, so that the corresponding BN nodes are binary. However, the BN framework is general and the extension to multi-state components is straightforward. Since the selected limit state is the least severe one (yield for bridges and slight/minor damage for unstable slopes), consequences of damage are limited to traffic reduction in terms of free-flow speed and capacity (closure is foreseen only for more severe damage states): in particular, it is assumed that for a damaged edge both properties reduce by 30%. All fragility curves used in this work are taken from the literature [23] and are lognormal cumulative distribution functions defined in terms of PGA (g). In particular, the fragility curve adopted for all 20 bridges has a median PGA of 0.12 g and $\sigma_{\log}$ = 0.44, while the one adopted for unstable slopes is characterized by a median PGA of 0.16 g and $\sigma_{\log}$ = 0.40, and a yield acceleration k_y = 0.05 g. The analysis is carried out at the traffic flow level, rather than purely in graph-theoretical terms of connectivity. For this purpose, an origin-destination (O-D) matrix, displayed in Table 2, is generated, with trips between the ten TAZs in vehicles per hour (vph). Such matrix is completely arbitrary, given that the scope of the paper is not the reliability analysis of the Pyrenees road network: however, the assumed trips are realistic for a pre-earthquake scenario. The pre-earthquake O-D matrix is used here given the illustrative character of the application, and also because, to the best knowledge of the authors, a methodology to establish reliably post-earthquake demands is one of the research gap in regional/urban seismic risk analysis, and it is obviously outside the scope of this paper. The interested reader is referred, e.g., to [24].

Table 2. Origin-destination matrix for the Pyrenees road network.

TAZ # From\To	1	2	3	4	5	6	7	8	9	10
1	0	200	500	300	150	500	200	250	350	250
2	200	0	500	300	150	500	200	250	350	250
3	300	300	0	250	120	600	150	200	200	150
4	350	300	200	0	120	80	150	100	250	180
5	60	60	50	200	0	40	50	40	90	60
6	300	250	450	300	80	0	100	80	350	160
7	300	300	200	300	80	300	0	40	80	100
8	300	300	200	300	80	300	40	0	80	100
9	400	400	250	500	100	400	100	50	0	120
10	150	150	100	100	40	160	80	160	80	0

The performance of the road network as a system is measured through two system performance measures, as detailed below. Both measures account for the capacity of the road edges to accommodate the traffic flow, and they may be either local (i.e., trip performance between two given TAZs) or global (i.e., aggregated trip performance over all TAZs):

1. *Global measure* S_1: Drivers' Delay (*DD*), defined as the difference between the congested (i.e., not free-flow) total travel time in damaged and normal, undamaged conditions (denoted with subscript "0"). Such total travel time is the sum of flow dependent edge travel times $TT(x)$ over all network edges, indexed by i, weighted by edge flows x:

$$DD = \sum_i x_i \times TT_i(x_i) - \sum_i x_{0,i} \times TT_{0,i}(x_{0,i}) \tag{5}$$

DD is a measure of lost passenger hours and was originally used as a factor to multiply a monetary value of the worked hour to give a proxy for indirect loss [25].

2. *Local measure* S_2: Local Drivers' Delay (*LDD*), which is simply the drivers' delay between two TAZs of interest with respect to normal conditions, due to damage suffered by the road network. This performance metric has the same definition as *DD*, Eq. (5), but with both summations extended over only the edges belonging to the shortest path between the two TAZs:

$$LDD\left(TAZ_{\#1}, TAZ_{\#2}\right) = \sum_{i \in \text{path}} x_i \times TT_i(x_i) - \sum_{i \in \text{path}} x_{0,i} \times TT_{0,i}(x_{0,i}) \tag{6}$$

In the case of interruption of all possible paths between the two TAZs, *LDD* is set to infinity. However, for the present application only the light damage limit state is considered for vulnerable edges, so that no breakage occurs and at least one path between all couples of TAZs always exists.

In both undamaged and damaged conditions, user equilibrium solved by the Frank-Wolfe algorithm is used to establish traffic flows and congested travel times on all network edges.

3.3. Monte Carlo Sampling and Initial Component Selection

Step 1 of the proposed approach consists in the generation of a dataset of the variables' states, by means of Monte Carlo sampling. To this end, the OOFIMS (Object Oriented Framework for Infrastructure Modelling and Simulation) platform [26] is used to model the road network and to sample 50,000 outcomes of the system's performance metrics, in terms of *DD* and *LDD*, according to the assumptions described in Section 3. Therefore, the OOFIMS platform outputs a state matrix of size [50,000 x 60], with the first 58 columns representing the components' states and the last two the performance metrics *LDD* and *DD*. This state matrix constitutes the dataset of descriptor/target variables for the creation of the random forest classification, from which unbiased importance measures are extracted in order to rank the components. As the random forest classification is specific to each system performance measure considered, two different sets of k=10 components are selected (this number is chosen here for illustration, based on the sensitivity to the number of components reported later on in 4.1): this amounts to 2^{10} = 1024 combinations of the parents' states for each variable S, which remains manageable in terms of computation cost. For each S, it is then possible to count the occurrences of the various combinations of the selected components' states, and to evaluate the conditional probabilities with Eq. (1). This process is exemplified in Table 3, displaying the 20 most frequent combinations of ten components, ranked by decreasing importance. The information reported in Table 3 reflects the component importance ranking. As an example, it can be noted that all the combinations involving damage for the first and most important component, C_{40} (see entries equal to 2 in the first column), result in zero occurrence of *DD* being in the first interval out of 50,000 samples (i.e., insignificant probability of occurrence). This observation indicates that damage on C_{40}

will severely affect the network's functionality: this is expected, given the importance of C_{40} in the estimation of *DD* (see also Figure 11b).

Table 3. Occurrences and probability estimation of *DD* being in the 1st discrete interval, for the 20 most frequent combinations of ten components, over the 50,000 outcomes of the state matrix. The ten components are ranked by decreasing importance.

ID	States (1=intact, 2=damaged) of the ten selected components										Total occurrences	Occurrences of the 1st *DD* interval	Estimated probability of the 1st *DD* interval
	C_{40}	C_{57}	C_{34}	C_{51}	C_{48}	C_{44}	C_{12}	C_{47}	C_{45}	C_{46}			
1	1	1	1	1	1	1	1	1	1	1	40835	40656	0.9956
2	1	1	1	1	1	1	2	1	1	1	743	699	0.9408
3	1	1	2	1	1	1	1	1	1	1	591	499	0.8443
4	2	1	1	1	1	1	1	1	1	1	587	0	0.0000
5	1	2	1	1	1	1	1	1	1	1	535	438	0.8187
6	1	1	1	1	2	1	1	1	1	1	431	403	0.9350
7	2	1	2	1	1	1	1	1	1	1	373	0	0.0000
8	1	1	1	1	1	1	1	2	1	1	343	272	0.7930
9	1	1	1	2	1	1	1	1	1	1	317	241	0.7603
10	1	1	1	2	1	1	1	2	1	1	260	57	0.2192
11	1	1	1	1	1	2	1	1	1	1	228	113	0.4956
12	1	1	1	1	1	1	1	1	1	2	196	74	0.3776
13	1	1	1	1	1	1	1	1	2	1	176	49	0.2784
14	2	1	2	1	1	1	2	1	1	1	158	0	0.0000
15	2	1	2	1	2	1	1	1	1	1	131	0	0.0000
16	2	1	1	1	1	1	2	1	1	1	121	0	0.0000
17	1	1	2	1	1	1	2	1	1	1	111	45	0.4054
18	2	1	2	1	2	1	2	1	1	1	89	0	0.0000
19	1	2	1	1	1	1	2	1	1	1	85	35	0.4118
20	2	1	1	1	2	1	1	1	1	1	78	0	0.0000
...	…	…	…	…	…	…	…	…	…	…	…	…	…

3.4. Bayesian Inference with the t-Naïve BN Formulation

Once the CPTs for both system metrics have been estimated from the state matrix, the *t-Naïve* BN is built by using an exact formulation down to the component nodes (i.e., as in Figure 1) and the approximate formulation from the component nodes to the S nodes. The resulting BN, composed of 355 nodes and 544 edges, is displayed in Figure 6. It can be seen that only ten edges converge to each S node: in particular, components C_{40}, C_{57}, C_{34}, C_{51}, C_{48}, C_{44}, C_{12}, C_{47}, C_{45} and C_{46} are linked to S_1

(*DD*), while components C_{40}, C_{34}, C_{46}, C_{44}, C_{45}, C_{48}, C_{55}, C_{12}, C_{54} and C_{52} are linked to S_2 (*LDD*). The components are ordered in descending order of importance ranking.

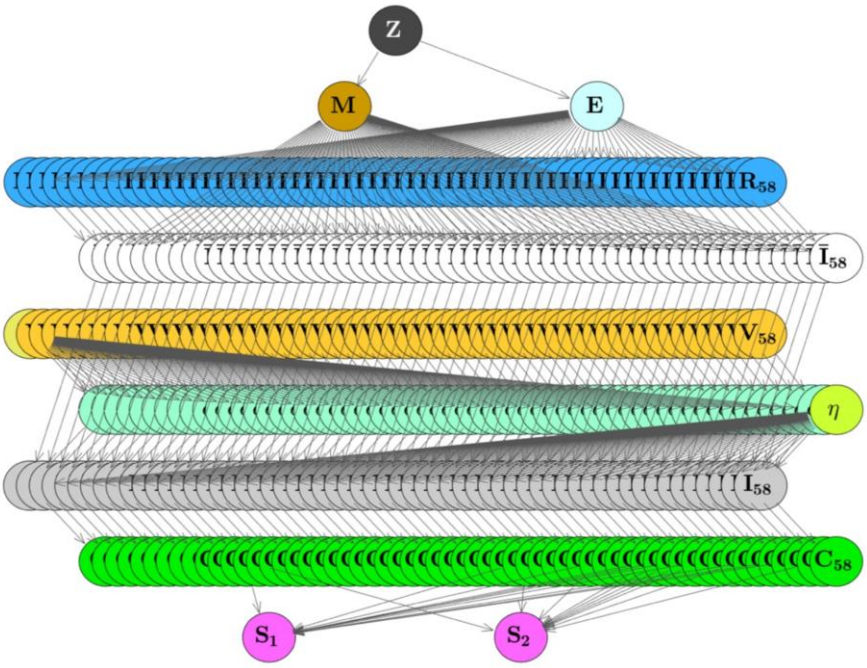

Figure 6. Layout of the t-Naïve BN formulation for the case-study, with ten components selected for each performance metric.

Figure 7 shows the location of the two sets of k=10 selected components, linked to the *LDD* and *DD* nodes, respectively. It can be noted that all the components related to *LDD* (red spots in the figure) are located in the shortest path between the TAZs of interest, #2 and #8, while the components linked to *DD* (blue circles in the figure) are scattered in the network, but anyway contained in the portion shown, which is therefore the most important one from the point of view of functionality.

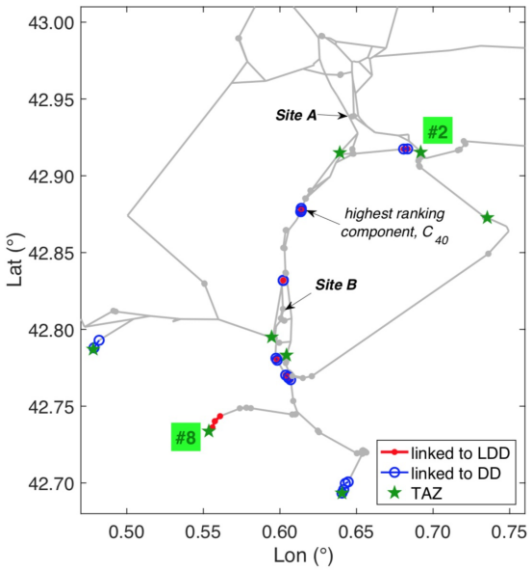

Figure 7. Zoomed-in area of the road network, showing the location of the ten selected components linked to the *LDD* (red) and *DD* (blue) nodes in the BN. Sites A and B represent vulnerable components that are used as a source of field observations in the Bayesian inference presented later.

This BN is implemented in the Bayes Net toolbox, BNT [27]. All continuous variables must be discretized beforehand, in order for exact inference engines such as the junction-tree algorithm to be used. The following discretization schemes are assumed for the continuous variables:

- Magnitude M: 10 intervals based on equal quantiles following the Gutenberg-Richter distribution (i.e., from wider intervals for low magnitudes to more refined intervals for large magnitudes), for each seismogenic zone;
- Epicenter location E: uniform intervals distributed among the seven seismogenic zones (see the green dots in Figure 8 corresponding to all discretized locations), for a total of 421 points;
- Epicentral distance R_i: 421 uniform intervals, based on the number of epicenter locations;
- GMPE error terms U, V_i, ε_i and η: 10 intervals based on equal quantiles following the standard normal distribution (i.e., wider intervals at the tails and narrower intervals around 0);
- Intensity measures $\bar{I}_i$ and I_i: 20 uniform intervals between the lowest and highest possible intensities;

- System performance measures S_1 and S_2: 10 uniform intervals between 0 (i.e., no loss) and the maximum sampled loss.

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. For instance, with the adopted discretization, the BN in Figure 6 leads to a junction tree with the largest clique size reaching a little more than 430,000,000 elements. Therefore this discretization process constitutes a potential source of errors and uncertainties, which will be discussed later in Section 4.3 along with potential refinement strategies.

Several inference operations are then 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 performance metrics S:

- 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 local performance metrics, 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 system metrics are practically unavailable in the case of road networks, only the observations at the level of the components are considered here. The inference scenarios on the BN are described in Table 4.

Table 4. Proposed inference scenarios for the demonstration of the BN applied to the road network. Sites A and B are shown in Figure 7 and the epicenter location in Figure 8.

Scenario ID	Evidence	Marginalized node
#0 (prior)	None	*LDD*, *DD*, I_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, I_A and I_B high	*LDD*, *DD*
#4	Epicenter (R_{avg} = 49 km), M_w 6.5, I_A high	I_B, C_B

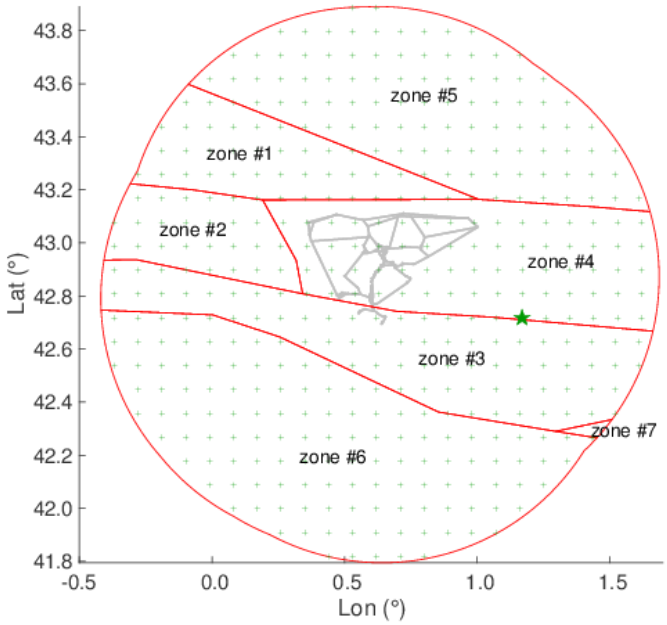

Figure 8. Location of the earthquake epicenter in inference scenarios #1 to #4 (green star), with respect to the seven seismogenic areas (red polygons).

Using a 2.40 GHz eight-core CPU with 32 GB RAM, the computational time is around 20 hours for the off-line Monte Carlo simulation (50,000 samples) and twelve minutes for the inference-related operations (i.e., component selection, BN creation and execution of all inference scenarios in Table 4, by the junction-tree algorithm). The prior and posterior distribution of all scenarios are detailed in Figure 9.

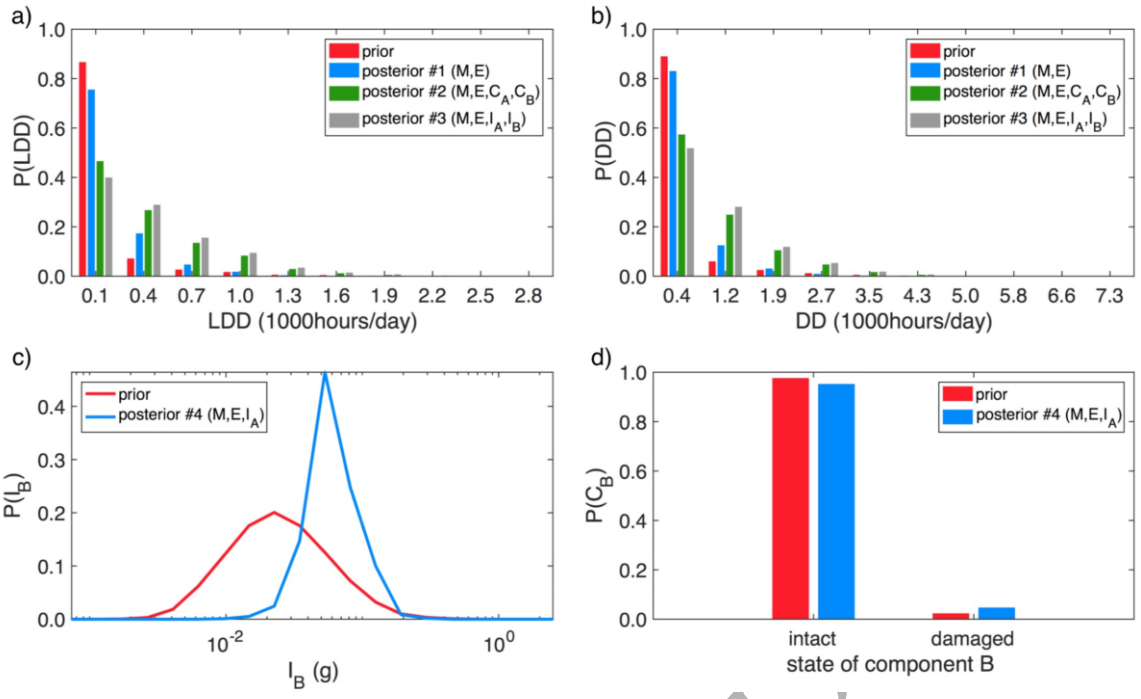

Figure 9. Prior and posterior distributions for the considered inference scenarios.

As shown in Figure 9, 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 metrics *DD* and *LDD*, evidence on their damage states or of the hazard intensity at their locations have 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: this effect is made possible by the statistical dependency between the I_i variables, which propagates the evidence to neighboring components and finally to the system performance metrics (e.g., see the two bottom plots in Figure 9).

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 metric 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.

Finally, it should be noted that the two bottom plots 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 I_i, which may lead to imprecise representations of the probability density functions. Inference scenario #4 appears to have a significant impact on the distribution of the hazard intensity at site B (I_B), however it does not lead to a huge change in the damage distribution of C_B: 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.

As already pointed out, evidence on damage states of not selected (i.e., excluded from the set) components has still an influence on the performance of the road network, due to the statistical dependency between the I_i variables. However, such influence can be only indirect, leading to possibly overlook important features derived from valuable pieces of information. In order to overcome this issue, the proposed methodology is adaptive and evidence-driven, in the sense that the components selected through an importance ranking algorithm can be replaced by other components for which evidence is available. In this way, all the information made available is directly used to update the posterior distribution of BN nodes. The replacement is undertaken in place of simple addition in order to keep the computational cost affordable, so that the number of components linked to S nodes remains unchanged. Figure 10 exemplifies this procedure, with reference to a base case with only five components linked to both system measures S. For this configuration, and using the same 2.40 GHz eight-core CPU with 32 GB RAM, the computational cost is 17 seconds for BN creation (including computation of CPTs for both S nodes) and just two seconds for a single marginalization on all 58 vulnerable components (i.e., estimation of prior or posterior distribution of component states given an evidence scenario). The importance ranking of components by random forest is carried out just once at

the beginning of the procedure, and for this case-study (i.e. 58 vulnerable components, m=2 system metrics S and a 50,000 sample Monte Carlo simulation) it takes six minutes.

The subplots in Figure 10 show the probability of failure of all 58 vulnerable components, with different color shades from white (low probability) to black (high probability), according to the following seven evidence scenarios.

a) No evidence (i.e., prior probability);

b) Epicenter (R_{avg} = 49 km east) and M_w 6.5;

c) Scenario b) plus damage evidence on the first and most important component (C_{40}) included in $\widetilde{\mathbf{C}}$, which is the vector (5×1) containing the IDs (ranging from 1 to 58) of the components selected through random forest;

d) Scenario c) plus damage evidence on a component not included initially in the set, which then replaces the fifth component of $\widetilde{\mathbf{C}}$;

e) f) and g) Scenario d) plus damage evidence on other three components not included initially in the set, which then replace the fourth, third and second components of $\widetilde{\mathbf{C}}$, respectively (according to the initial numbering).

The subplot b) shows that, given evidence on M and E located 49 km to the East, the two easternmost components have an increase in failure probability, as expected. Subplots from c) to g) clearly display the failure probability increasing over the network, as new evidence on component damage becomes available. It is important to note that the introduction of evidence on the fourth and fifth components (see subplots f) and g)) does not have a large impact on probabilities across the network, with only a few components reaching the last probability range (0.48-0.57, see Figure 10 legend) and thus possibly changing their mode (i.e., the damaged state becomes the most likely). This suggests that the methodology is able to give emergency managers a clear indication of updated failure probability across a realistic network with just a few pieces of evidence, in near-real time.

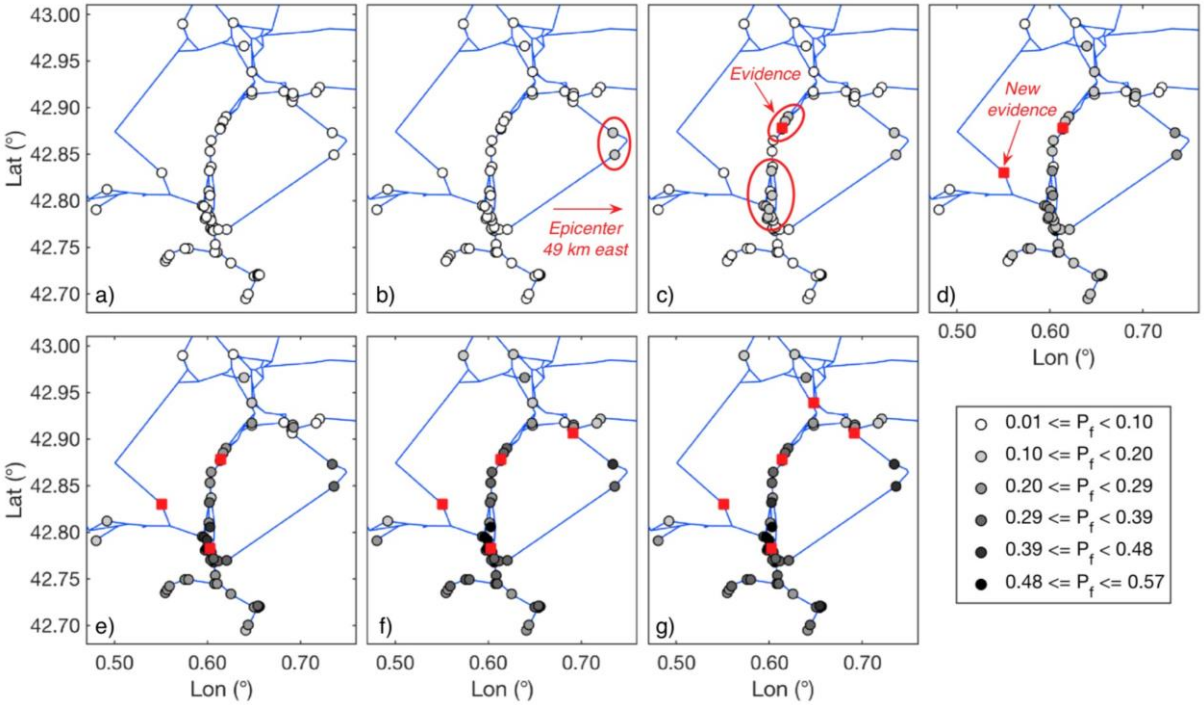

Figure 10. Probability of failure of components according to: a) the prior distribution and b) to g) the posterior distributions for the considered successive inference scenarios. The evidenced components are shown with red squares.

4. Sensitivity Analysis

In order to be guided in the choice of the number of selected components, the number of simulation samples and the discrete intervals of some continuous parameters, several sensitivity analyses have been carried out. In this section, the results of such analyses are presented, serving also the purpose to justify the assumptions made in the first part of the paper.

4.1. Components Selection

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 4) of *DD*, in particular the first *DD* state; then the statistics (mean and standard deviation) of the inference results have been computed. Table 5 highlights that the highest value of standard deviation is still very low, thus confirming the robustness of the component selection via random forest.

Table 5. 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	0.8821	0.8252	0.5604
St. Dev.	0.0016	0.0064	0.0099

Finally, in order to investigate the sensitivity of the solution to the number of components retained in the *t-Naïve* formulation, the prior and two posterior distributions (#1 and #2 in Table 4) of *DD* have been computed with a variable number of components linked to the *DD* node, from one to thirteen. Figure 11a), presents the results with reference to the first *DD* state. Taking the values with thirteen components as "exact", it can be seen that considering more than seven or eight components does not practically change the probability values of the first *DD* state: it is thus possible to conclude that the performance is quite well captured with even less than ten components. This is also evidenced by Figure 11b), where the normalized unbiased importance measure reaches larger values for the first few components (with some clear gaps between them) and attains quite low values after the tenth one (i.e., C_{46}). This clearly indicates that the remaining components do not play an important role in the estimation of the quantities of interest. Based on these results, the number of components for the inferences of Figure 9 has been set to ten, which is a good compromise between accuracy in the results and computational effort.

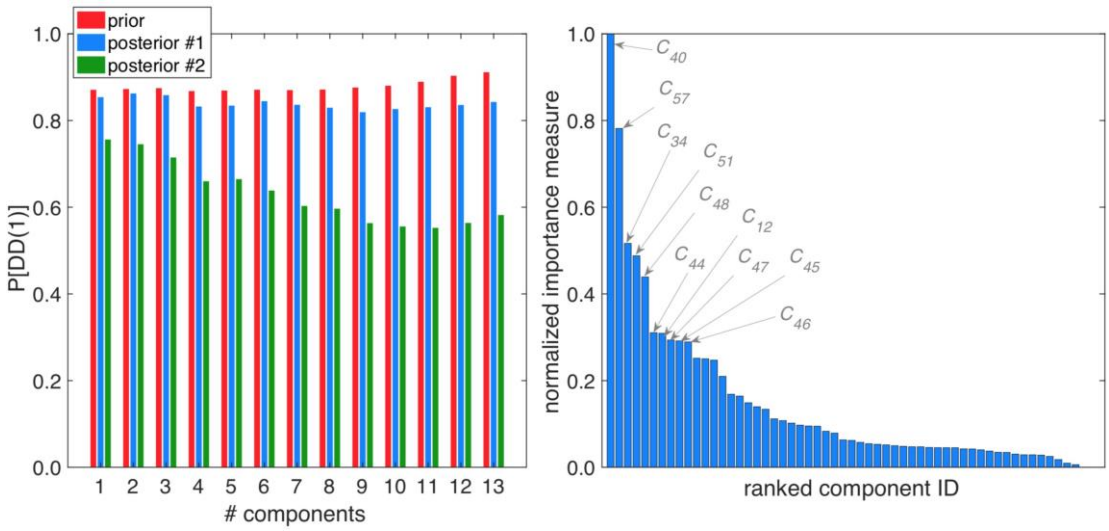

Figure 11. a) Sensitivity of the prior and two posterior distributions of Drivers' Delay to the number of selected components, and b) normalized importance measure of the 58 components ranked by decreasing importance.

As indicated above, all inferences in this work have been carried out via junction-tree algorithm. With this exact inference engine and the *t-Naïve* formulation, and given the hardware properties of the employed computer, the highest number of components that it is possible to retain to get the solution in a reasonable amount of time resulted to be thirteen (see Figure 11b)). In order to exceed this number and get a more accurate solution, a possible choice would be changing the inference engine and in particular trying an approximate inference. Among the possible approximate engines available in the BNT, the authors found that the likelihood weighting algorithm [28], which performs an importance sampling where the weights are based on the likelihood of the evidence; is the only feasible one. However, the inference analyses took long computational times and yielded very unstable results, which are not worth the presentation in this paper. More robust sampling-based inference algorithms, such as Gibbs sampling, deserve further investigations, although they are currently not implemented in the BNT.

It may be argued that 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. However, 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.

4.2. Number of Simulation Samples

As already said, the analyses in Section 3 are based on a Monte Carlo simulation with 50,000 samples, carried out in the OOFIMS platform and producing a state matrix of size [50,000×60]. In order to get an increasingly more refined solution, tending to the exact one, one option is simply increasing the number of samples. This will of course require more computational time. However, it is noted that the simulation is carried out off-line, before the occurrence of an earthquake event, and therefore an increase in computational time does not compromise the capability of BNs to be used as rapid response systems, able to update damage and loss predictions in near-real time from field observations. In order to gain more insight into the increase of accuracy with the number of samples, bootstrap analysis has been performed starting from the initial 50,000 samples. Employing bootstrap, the vectors of normalized importance measure of all 58 components, as derived from the random forest classification for *DD*, as well as the probability of the first *DD* state, according to the prior and two posterior distributions (#1 and #2 in Table 4), have been re-evaluated on the base of randomly drawn sub-samples of the larger state matrix with 50,000 rows. For this exploration, 100 random samples have thus been drawn for each sample size *K* from 10,000 to 40,000 with steps of 10,000. Figure 12a) shows the frequency histograms of a measure of similarity between the generic importance vector, derived from a random sub-sample of the state matrix, and the reference importance vector obtained from the entire state matrix of size 50,000. Calling $\boldsymbol{\Theta}$ the generic importance vector (58×1) containing the normalized unbiased importance measure (ranging between 0 and 1) of components, and $\boldsymbol{\Theta}_{\text{ref}}$ the reference importance vector, the proposed similarity measure, ranging between 0 and 1, is then defined as:

$$\text{Similarity} = \frac{\boldsymbol{\Theta}^T \cdot \boldsymbol{\Theta}_{\text{ref}}}{\left\|\boldsymbol{\Theta}\right\| \left\|\boldsymbol{\Theta}_{\text{ref}}\right\|} \tag{7}$$

The frequency histograms in Figure 12a) are based on 100 values of similarity for each sample size. It is possible to see that, taking as "exact" the results with 50,000 samples, there is a quite small reduction in similarity, meaning that the sub-sample simulations lead to very similar importance vectors, especially for sample sizes from 20,000 up; the histograms for 30,000 and 40,000 samples appear to be almost superimposed, with similarity values approaching unity. The sensitivity of probability distributions to the sample size is analyzed in Figure 12b), displaying the 5% and 95% fractile of the distribution of P[*DD*(1)] (according to the prior and two posterior distributions) normalized by the reference value (i.e., related to 50,000 samples). Figure 12b) shows that the goodness of results is linked to the type of distribution one is interested in: in particular, the prior distribution is practically insensitive to the sample size, while for the posterior #2 one can expect a divergence up to 15% for 10,000 samples, which could be even considered acceptable for such a cheap analysis. Again, especially for sample sizes from 20,000 up, the sensitivity to sample size is quite low: as an example, it is possible to state that, with 90% probability, the posterior #1 will fall within [+1,+4%] of the reference value for K as low as 20,000, and within [-1,+3%] for K = 30,000. These findings make the overall approach robust against the user-defined number of simulation runs.

Finally, coming back to the initial issue of this section, another possible option to obtain a more refined solution is using "variance reduction techniques", such as importance sampling, which increase the accuracy of results at parity of simulation runs, or allow to get a fixed accuracy carrying out a lower number of samples. This issue has not been investigated herein. Future work will investigate the feasibility of adopting an importance sampling scheme in place of plain MC to inform the BN.

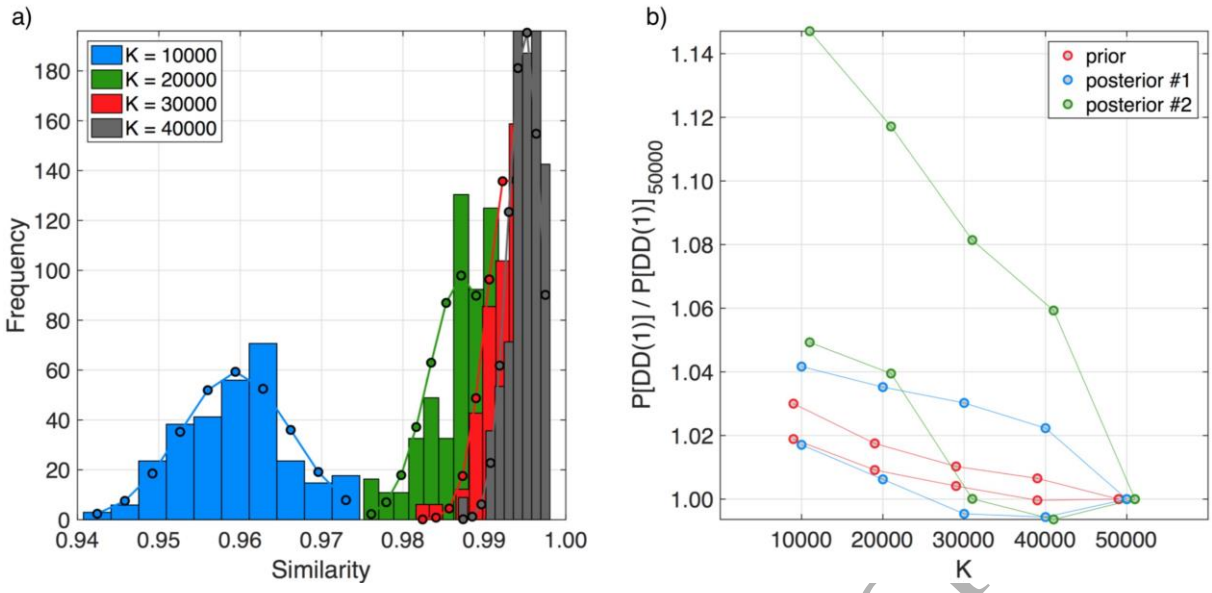

Figure 12. Bootstrap results, based on 100 random samples of size *K*. a) Frequency histograms, with normal distribution fit superimposed, of Similarity for different sample sizes; b) curves of 5% and 95% fractile of P[*DD*(1)], according to the prior and two posterior distributions, normalized by the reference value.

4.3. Variable discretization

As pointed out above, in the present case most of the variables are continuous and have been discretized beforehand, in order to use the exact inference. It is important to carefully evaluate the consequences of the discretization, as suggested by Nojavan et al. [29]. Their work identified three commonly used discretization methods, each designed to capture certain features of the data distribution, namely *(i)* Equal interval, *(ii)* Equal quantile and *(iii)* Moment matching. The sensitivity of the inference results to the discretization adopted has been investigated also in this work. First of all, the focus has been put on the root nodes of the considered BN, namely M, U, V_i and η. Since the nodes U, V_i and η have a normal distribution, they have been discretized with the equal quantile method, so to capture their mode. On the other hand, M follows the Gutenberg-Richter distribution and has been discretized with equal intervals in the original version of the formulation. This discretization is not adapted to the problem at hand, i.e. the identification of failure events in the infrastructure: in fact, a more refined discretization is needed over the portions of the range associated

with high probability mass [3]. Therefore, also for the magnitude node the equal quantile method (called here "smart discretization") has been applied and its influence on the inference results has been evaluated. Table 6 summarizes the sensitivity results, in terms of influence of the smart discretization and number of states (5, 10 or 20) of *M* on the prior and two posterior distributions (#1: Epicenter only, #2: Epicenter+C_A+C_B) of the first *DD* state. It can be noted that the higher the number of states, the larger the influence of the more accurate discretization on the inference results. Since the highest magnitudes are better captured, the probabilities of the first *DD* state tend to decrease, with consequent increase of the probabilities at higher states: this means that the predicted impact on the performance metric is larger, as expected. For the inferences of Figure 9, the smart discretization for the magnitude node with ten states has been adopted.

Table 6. Sensitivity of the inference results to the discretization (type and number of states) of the magnitude node. The results are referred to the probability of the first state of *DD*, according to the prior distribution and two posterior distributions.

Smart discretization	# states	Inference type		
		$P[DD(1)]_{prior}$	$P[DD(1)]_{posterior\#1}$	$P[DD(1)]_{posterior\#2}$
NO	5	0.9084	0.9651	0.5374
YES	5	0.9008	0.9501	0.5296
NO	10	0.9024	0.9611	0.5653
YES	10	0.8803	0.9310	0.5498
NO	20	0.8996	0.9592	0.5831
YES	20	0.8666	0.9205	0.5513

Coming to BN nodes that are not roots, a smart discretization cannot be applied a priori, given the lack of a marginal distribution. For such nodes it is first needed to enter evidence and retrieve a posterior distribution. Based on the latter, it is possible to refine the discretization iteratively, again with the aim to better capture the distribution over the portions of the range associated with high probability mass. To test the feasibility and the influence of a smart discretization on target variables, in this work the *DD* node's distribution has been refined following an iterative approach, along the lines of the one proposed by Neil et al. [30], named *dynamic discretization*. Firstly, the target node's distribution is

initialized, following in our case the equal interval method. Then evidence is entered and the inference is performed, retrieving the target node's posterior distribution. A new discretization is then created, by splitting in two halves the interval with the highest probability and merging two consecutive intervals with the lowest probability. The process is repeated a user-defined number of times and is stopped if the lowest probability in the current distribution is higher than a user-defined threshold. Figure 13 shows the effects of the dynamic discretization, with five iterations, on the *DD* node's distribution, using the posterior distribution given scenario #2 in Table 4 as target distribution. The threshold has been set to 0.01, meaning that only if the lowest probability is less than 0.01 another step will be undertaken. The focus in the figure is on the first four states, the most significant for this variable. It has been seen how the process leads to increase the number of intervals (eight in place of four) in the same *DD* range (0-3000 hours/day), thus better capturing not only the posterior #2 but also the prior and posterior #1. In the subplots of Figure 13a) and Figure 13b), the *DD* range has been limited to the most important intervals, in order to improve the figure's readability and thus better highlight the interval increase. Based on these results, the subplots in Figure 9 related to the performance metrics *LDD* and *DD* could be updated a user-defined number of times: this is not shown here since the aim of Figure 9 is just to show the capabilities of the framework with reference to a few sample inference scenarios. It should be noted that this smart discretization scheme may only be carried out after an earthquake event has occurred, since it is based on the posterior distribution once evidence has been entered. Depending on the time taken to perform one inference, however, several iterations of the discrete intervals might still be feasible in order to deliver a refined loss distribution within the imposed timeframe. Still, the first iteration without any smart discretization may be delivered as a first approximation, before more refined distributions are generated. Another possible approach might be supervised discretization, which scores all possible discretization schemes, based on the same principle as supervised BN learning.

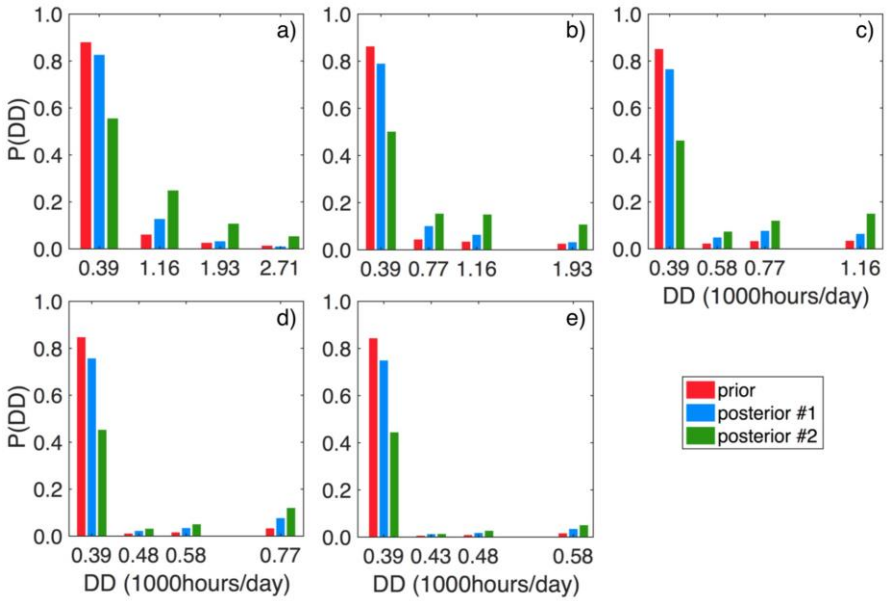

Figure 13. Dynamic discretization of the target variable *DD* (five iterations), using the posterior distribution given scenario #2.

5. Conclusions

Starting from current computational and conceptual challenges regarding the use of BNs for the seismic loss analysis of spatially distributed infrastructure systems, this paper has detailed a three-step approach that builds a simplified BN structure based on a preliminary off-line Monte Carlo simulation of the system. By approximating the probabilistic relation between the state of the components and the distribution of a system performance measure of interest, it has been shown that a naïve BN formulation (i.e., converging structure from the components to the system node) is able to provide satisfying probability estimates, even when considering a fraction of the vulnerable components. This encouraging result is due to two main factors:

(i) 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;

(ii) the BN adopts (and slightly modifies) the exact structure by Bensi et al. [3] for all variables in the seismic hazard portion of the BN, and down to the component states, while only the system portion is approximate.

The application of this so-called *t-Naïve* BN formulation to a real-world road network in France has demonstrated its potential in a backward analysis framework, when a Bayesian inference is produced from field evidence such as the recording of ground-motion intensities or the observation of damaged components at various locations. For this specific case-study, stable posterior probability estimates could be obtained even with a handful of selected components. 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. The extension of this approach to other hazards such as floods or landslides depends strongly on the type of hazard propagation models used, since it has been shown here that the spatial correlation of the hazard intensity is one of the main drivers of the statistical dependence between the component damage events. 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: it has been shown that the uncertainty introduced by this issue, which is often overlooked, may be comparable to the one due to the approximate BN formulation.

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.

References

[1] Erdik M, Şeşetyan K, Demircioğlu MB, Hancilar U, Zülfikar C. Rapid earthquake loss assessment after damaging earthquakes. Soil Dyn Earthq Eng 2011;31:247-266.

[2] Bensi M, Der Kiureghian A, Straub D. Framework for post-earthquake risk assessment and decision making for infrastructure systems. ASCE-ASME J Risk Uncertainty Eng Syst, Part A: Civ Eng 2015;1:1-17.

[3] Bensi M, Der Kiureghian A, Straub D. Bayesian network modeling of correlated random variables drawn from a Gaussian random field. Struct Saf 2011;33:317-332.

[4] Bensi M, Der Kiureghian A, Straub D. Efficient Bayesian network modeling of systems. Reliab Eng Syst Saf 2013;112:200-213.

[5] Cavalieri F, Franchin P, Gehl P, D'Ayala D. Bayesian networks and infrastructure systems: Computational and methodological challenges. In: Gardoni P, editor. Risk and reliability analysis: Theory and applications. Springer; 2017. p. 385-415.

[6] Tien I, Der Kiureghian A. Algorithms for Bayesian network modeling and reliability assessment of infrastructure systems. Reliab Eng Syst Saf 2016;156:134-147.

[7] Tien I, Der Kiureghian A. Reliability Assessment of critical infrastructure using Bayesian networks. J Infrastruct Syst 2017;23.

[8] Pozzi M, Der Kiureghian A. Gaussian Bayesian network for reliability analysis of a system of bridges. Proceedings of the 11th International Conference on Structural Safety and Reliability; 2013 Jun 16-20; New York, United States.

[9] Hosseini S, Barker K. Modeling infrastructure resilience using Bayesian networks: A case study of inland waterway ports. Comput Ind Eng 2016;93:252-266.

[10] Hong L, Ouyang M, Peeta S, He X, Yan Y. Vulnerability assessment and mitigation for the Chinese railway system under floods. Reliab Eng Syst Saf 2015;137:58-68.

[11] Cavalieri F, Franchin P, Buriticá Cortés JA, Tesfamariam S. Models for seismic vulnerability analysis of power networks: comparative assessment. Comput-Aided Civ Infrastruct Eng 2014;29:590-607.

[12] Breiman L. Random Forests. Mach Learn 2001;45:5-32.

[13] Woessner J, Danciu L, Kaestli P, Monelli D. Database of seismogenic zones, Mmax, earthquake activity rates, ground motion attenuation relations and associated logic trees. FP7 SHARE Project Deliverable D6.6; 2013.

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

[15] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[16] MATLAB and Statistics Toolbox Release 2013a, The MathWorks, Inc., Natick, Massachusetts, United States.

[17] Wei P, Lu Z, Song J. Variable importance analysis: a comprehensive review. Reliab Eng Syst Saf 2015;142:399–432.

[18] Doguc O, Ramirez-Marquez JE. A generic method for estimating system reliability using Bayesian networks. Reliab Eng Syst Saf 2009;94:542–550.

[19] Zwirglmaier K (2016). Reliability analysis with Bayesian networks [PhD dissertation]. Munich (Germany): Technische Universität München; 2016.

[20] Gehl P, Cavalieri F, Franchin P, Negulescu C. 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; 2017 Aug 6-10; Vienna, Austria.

[21] Akkar S, Bommer JJ. Empirical equations for the prediction of PGA, PGV, and spectral accelerations in Europe, the Mediterranean region, and the Middle East. Seismol Res Lett 2010;81:195-206.

[22] Bommer JJ, Akkar S, Drouet S. Extending ground-motion prediction equations for spectral accelerations to higher response frequencies. Bull Earthq Eng 2012;10:379-399.

[23] Kaynia AM., Taucer F, Hancilar U. Guidelines for deriving seismic fragility functions of elements at risk: buildings, lifelines, transportation networks and critical facilities. FP7 SYNER-G Project Reference Report 4, Publications Office of the European Union; 2013. ISBN 978-92-79-28966-8.

[24] Cho S, Murachi Y, Fan Y, Shinozuka M. Transportation network simulation for dynamic origin-destination matrix under earthquake damage. Proceedings of the 13th World Conference on Earthquake Engineering; 2004 Aug 1-6; Vancouver, Canada.

[25] Shinozuka M, Murachi Y, Dong X, Zhou Y, Orlikowski MJ. 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; 2003 Feb 21-22; Beijing, China.

[26] Franchin P, Cavalieri F. OOFIMS, Object-Oriented Framework for Infrastructure Modeling and Simulation [Accessed 2018 Feb 6]. Available from: https://sites.google.com/a/uniroma1.it/oofims/.

[27] Murphy K. The Bayes Net toolbox for Matlab. Comput Sci Stat 2001;33:1024-1034.

[28] Fung R., Chang K, 1990. Weighting and integrating evidence for stochastic simulation in Bayesian networks. In: Henrion M, Shachter R, Kanal L, Lemmer J, editors. Uncertainty in artificial intelligence. North-Holland, Amsterdam; 1990. p. 209–220.

[29] Nojavan F, Qian SS, Stow CA. Comparative analysis of discretization methods in Bayesian networks. Environ Model Softw 2017;87:64-71.

[30] Neil M, Tailor M, Marquez D. Inference in hybrid Bayesian networks using dynamic discretization. Stat Comput 2007;17:219-233.

GraphicalAbstract