Decision making and modelling uncertainty for the multi-criteria analysis of complex energy systems

Tairan Wang

CentraleSupélec

THÈSE

présentée par

Tairan WANG

pour l'obtention du

GRADE DE DOCTEUR

Spécialité : Génie Industriel

Laboratoire d'accueil : Laboratoire de Génie Industriel

SUJET :

Decision Making and Modelling Uncertainty for the Multi-criteria Analysis of Complex Energy Systems

La Prise de Décision et la Modélisation d'Incertitude pour l'Analyse Multi-critère des Systèmes Complexes Énergétiques

soutenue le : 8 juillet 2015

devant un jury composé de :

Ahti SALO	**Aalto University**	**Reviewer**
Vytis KOPUSTINSKAS	**European Commission, Joint Research Center**	**Reviewer**
François BEAUDOUIN	**Électricité de France R&D**	**Examiner**
Maria Francesca MILAZZO	**Università degli Studi di Messina**	**Examiner**
Enrico ZIO	**CentraleSupélec**	**Supervisor**
Vincent MOUSSEAU	**CentraleSupélec**	**Co-Supervisor**

2015ECAP0036

Acknowledgements

I would like to express my sincere appreciation to my thesis supervisor, Professor Enrico Zio, for picking me as his Ph.D. student at the first place. I am so lucky and honored to be supervised by a talented researcher, delegent professor like him. After these three years working with him, he is now far more than that but an uncle to me. All the weekly presentations of working progress, the discussions, the talks on the way to the metro station (after which he still had a flight to catch) had guided me to a promising path of my research and encouraged me to think and work even harder to finally conquer the new chanllenges. His strong enthusiasm and passion for scientific research in different domains, his prudent attitude and rigorous organization of the work had set a great career role model for me. As his student, you can actually feel his care. Thanks to all these, when I had difficulties on work, when I was ill or when I had emotional problems, I didn't give up but fought more bravely. His patience and comprehension is always a great support.

I would like to then express my sincere appreciation to my thesis co-supervisor, Professor Vincent Mousseau. Based on the characteristics of my thesis subject which combines two different scientific domains (risk analysis and decision making), it was him who had shown me the essence of the second aspect. There were many times that we spent several hours together discussing freely and equally how to build the model etc. Even though, sometimes, the two professors had different perspectives and demands on my research, the fact that I worked with two professors at the same time is definitely a plus without any conflicts.

I would like to extend my sincere gratitude to my co-adviser, Doctor Nicola Pedroni for being always present and available to discuss and solve all different kinds of problems with me. His insights, advices and patience had helped me to overcome the difficulties along the way. He had dedicated a lot of energy on my study, especially on this final thesis. It has been efficient, harmonious and unforgettable to work with

him. Without him, I could not have accomplished so well this work.

My deep appreciation goes to all the jury members, Professor Ahti Salo, Professor Vytis Kopustinskas, Doctor François Beaudouin and Doctor Maria Francesca Milazzo. I am very pleased to have presented my Ph. D. work in front of such a high qualified jury and have had such an interesting and fruitful Question & Answer section. Special thanks to the reviewers, Professors Ahti Salo and Professor Vytis Kopustinskas, for the evaluation of the manuscript and all the constructive and helpful remarks.

I would like to acknowledge Professor Jean-Claude Bocquet, ex-Director of the Industrial Engineering Laboratory (LGI), for a three-year hosting: it has been a pleasant experience working in this laboratory.

My warmest thanks go to Corinne Ollivier, Delphine Martin and Sylvie Guillemain, the secretaries of LGI. They have been so kind and lovely, especially to a forgein student as me, they have been so helpful both during the work and also the daily life. I would like to thank all of my colleagues from LGI. It's my great pleasure to be working here, to be one of the members of the family. And especially, I want to thank every chair member: Yanfu, Elisaveta, Ronay, Rodrigo, Yiping, Jie, Lo, Elisa, Yanhui, Valeria, Carlos, Ionela, Xing, Muxia, Fangyuan, Shanshan and Pietro. I am so lucky to work with these wonderful people! We are more than collegues. Besides, I would like to give my special thanks to my officemate, dear Elisa, Yiping, Fangyuan and Pietro. It is a funny, fruitful and unforgettable memory to work with them together in the same office. I even learnt to speak italian which made them laugh a lot.

During these three years of study in France, there are a lot of people outside school that have always been there for me which enriched my life and made France the second homeland for me: Dr. Nicole Bru, my "marraine" for already ten years, it was thanks to her that I began all my story with France in Ecole Centrale de Pékin; Dear Michèle Roche, my french aunt, who had shown me and taught me a lot about french

art, history and culture; Dear Barbara Simon, my french mom who has been taking care of me since the first day I stepped on France seven years ago; Dear Françoise & Jean-Claude Allanche, my french grandparents, who have had discussions with me as a schoolfellow, who have guided me as a teacher and who have treated me as a family member. Thanks to all these people, my life in France is more colorful, meaningful and peaceful.

Finally, I want to dedicate this thesis to my parents who have done so much to raise me up, to educate me, to guide me under all circumstances. Their expectation, their turst, their support, their tolerance and their love will always be the motivation for me to dream bigger and try harder all through my life.

Abstract:

This Ph. D. work addresses the vulnerability analysis of safety-critical systems (e.g., nuclear power plants) within a framework that combines the disciplines of risk analysis and multi-criteria decision-making.

The scientific contribution follows four directions: (i) a quantitative hierarchical model is developed to characterize the susceptibility of safety-critical systems to multiple types of hazard, within the needed 'all-hazard' view of the problem currently emerging in the risk analysis field; (ii) the quantitative assessment of vulnerability is tackled by an empirical classification framework: to this aim, a model, relying on the Majority Rule Sorting (MR-Sort) Method, typically used in the decision analysis field, is built on the basis of a (limited-size) set of data representing (a priori-known) vulnerability classification examples; (iii) three different approaches (namely, a model-retrieval-based method, the Bootstrap method and the leave-one-out cross-validation technique) are developed and applied to provide a quantitative assessment of the performance of the classification model (in terms of accuracy and confidence in the assignments), accounting for the uncertainty introduced into the analysis by the empirical construction of the vulnerability model; (iv) on the basis of the models developed, an inverse classification problem is solved to identify a set of protective actions which effectively reduce the level of vulnerability of the critical system under consideration. Two approaches are developed to this aim: the former is based on a novel sensitivity indicator, the latter on optimization.

Applications on fictitious and real case studies in the nuclear power plant risk field demonstrate the effectiveness of the proposed methodology.

Keywords: Safety-critical system, vulnerability analysis, risk analysis, multi-criteria decision making, malevolent intentional attacks, hierarchical model, Majority Rule Sorting (MR-Sort) classification model, Monte Carlo simulation, model-retrieval based approach, bootstrap method, cross-validation, classification accuracy, confidence estimation, protective actions, inverse classification problem, sensitivity analysis, nuclear power plants

Résumé:

Ce travail de thèse doctorale traite l'analyse de la vulnérabilité des systèmes critiques pour la sécurité (par exemple, les centrales nucléaires) dans un cadre qui combine les disciplines de l'analyse des risques et de la prise de décision de multi-critères.

La contribution scientifique suit quatre directions: (i) un modèle hiérarchique et quantitative est développé pour caractériser la susceptibilité des systèmes critiques pour la sécurité à plusieurs types de danger, en ayant la vue de 'tous risques' sur le problème actuellement émergeant dans le domaine de l'analyse des risques; (ii) l'évaluation quantitative de la vulnérabilité est abordé par un cadre de classification empirique: à cette fin, un modèle, en se fondant sur la Majority Rule Sorting (MR-Sort) Méthode, généralement utilisés dans le domaine de la prise de décision, est construit sur la base d'un ensemble de données (en taille limitée) représentant (a priori connu) des exemples de classification de vulnérabilité; (iii) trois approches différentes (à savoir, une model-retrieval-based méthode, la méthode Bootstrap et la technique de validation croisée leave-one-out) sont élaborées et appliquées pour fournir une évaluation quantitative de la performance du modèle de classification (en termes de précision et de confiance dans les classifications), ce qui représente l'incertitude introduite dans l'analyse par la construction empirique du modèle de la vulnérabilité; (iv) basé sur des modèles développés, un problème de classification inverse est résolu à identifier un ensemble de mesures de protection qui réduisent efficacement le niveau de vulnérabilité du système critique à l'étude. Deux approches sont développées dans cet objectif: le premier est basé sur un nouvel indicateur de sensibilité, ce dernier sur l'optimisation.

Les applications sur des études de cas fictifs et réels dans le domaine des risques de centrales nucléaires démontrent l'efficacité de la méthode proposée.

Mots-clés: système critique pour la sécurité, analyse de la vulnérabilité, analyse des risques, prise de décision de multi-critères, attaques intentionnelles et malveillantes, modèle hiérarchique, modèle de classification basé sur Majority Rule Sorting (MR-Sort), simulation Monte Carlo, model-retrieval-based méthode, méthode Bootstrap,

validation croisée, précision de la classification, estimation de la confiance, mesures de protection, problème de classification inverse, analyse de sensibilité, centrales nucléaires

Contents

List of Figures

List of Tables

Acronyms

CI Critical Infrastructure

NSF National Science Foundation

NPP Nuclear Power Plant

LGI Laboratoire Génie Industriel

EDF Électricité de France

MCDM Multi-Criteria Decision Making

MCDA Multi-Criteria Decision Aid

MAVT Multi-Attribute Value Theory

MAUT Multi-Attribute Utility Theory

ACUTA Analytic Centre UTilité Additive

ELECTRE ÉLimination Et Choix Traduisant la REalité

LAMSADE Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision

PROMETHEE Preference Ranking Organization Method for Enrichment and Evaluation

MR-Sort Majority Rule Sorting

GTST Goal Tree Success Tree

MLD Master Logic Diagram

MFM Multilevel Flow Modelling

GTST-DMLD GTST–Dynamic MLD

PDA Preference Disaggregation Analysis

UTA UTilité Additive

DM Decision Maker

ORT Outranking Relation Theory

MIP Mixed Integer Program

LOOCV Leave-One-Out Cross-Validation

Appended papers

Paper (i)

T. R. Wang, V. Mousseau, and E. Zio. A Hierarchical Decision Making Framework for Vulnerability Analysis. European Safety and RELiability Conference (ESREL2013), Sep 2013, Amsterdam, Netherlands. pp.1-8.

Paper (ii)

T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio. Assessing the performance of a classification-based vulnerability analysis model, Risk Analysis, Vol.35, No.9, 2015.

Paper (iii)

T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio. An empirical classification-based framework for the safety-related criticality assessment of complex energy production systems, in presence of inconsistent data, Reliability Engineering & System Safety, 2015, under review.

Paper (iv)

T. R. Wang, N. Pedroni, and E. Zio. Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent intentional acts: a sensitivity-based decision-making approach, Reliability Engineering & System Safety, 2015, accepted.

Paper (v)

T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio. Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent intentional acts: an optimization-based decision-making approach, European Journal of Operational Research, 2015, submitted.

Chapter 1

INTRODUCTION

The focus of the present Ph.D. thesis is on the qualitative and quantitative assessment and management of the vulnerbility of safety-critical systems to intentional hazards (i.e., malevolent attacks). In the work developed, the vulnerability to intentional hazards is first analyzed and represented within a hierarchical framework that systematically decompose it into the factors which influence it, down to "basic" factors for which data and information can be collected. Then, the vulnerability analysis is done qualitatively and quantitatively: first, a ranking method is employed to obtain a "comparative" (relative) evaluation of the vulnerability of a group of safety-critical systems; then, an empirical classification model is used to provide a quantitative (absolute) assessment of system vulnerability. The accuracy and confidence of the vulnerability assignments are also estimated, to cope with the uncertainty intrinsic in the classification model. Finally, vulnerability is managed by optimizing the strategy of protective actions to reduce it, based on sensitivity indicators and optimization methods. The applications considered in this thesis work regard energy systems and, in particular, nuclear power plants-NPPs. The work has been performed at the Laboratoire Génie Industriel (LGI, Industrial Engineering Laboratory) under the Chair on Systems Science and the Energetic Challenge, Foundation EDF, at Centrale & Supélec, Paris, France.

This chapter aims to provide a general overview of the problems addressed in this dissertation, and is organized as follows. The concept of vulnerability is defined and its

characteristics are introduced in Section 1.1; in Section 1.2 the multi-criteria decision making framework of vulnerability is discussed; Section 1.3 illustrates the issues and the motivation of the research conducted; the synthesis of the contributions of the thesis is drawn in Section 1.4; finally, in Section 1.5, the structure of the dissertation is given.

1.1 Vulnerability

While the concept of risk is fairly mature and consensually agreed on, the concept of vulnerability is still evolving and not yet established [76]. Risk is considered to be quantifiable in terms of the probability of occurrence (frequency) of a specific (mostly undesired/adverse) event leading to loss damage or injury, and its extent. Vulnerability has been introduced to give an hazard-centric perception of disasters, which would be too limited in terms of risks. The claim is that the estimated failure probabilities quantified in a risk assessment and used in risk management to inform decisions may be unreliable, if marred by insufficient knowledge and inappropriate assumptions; moreover, unexpected events can occur, like unknown failure mechanisms [61]. To inform the risk scenario one must, then, consider the level of vulnerability, which makes the difference between a hazard of low intensity that could have severe consequences and a hazard of high intensity that could have negligible consequences [128].

Two main interpretations of vulnerability have been given: one relative to a global system property and another one quantifying directly system components. The first interpretation (closer to the concept of risk), seeks to account for the extent of adverse effects caused by the occurrence of a specific hazardous event [61][76].

The concept of vulnerability as global system property embeds three other concepts [76]:

- degree of loss and damages due to the impact of a hazard;
- degree of exposure to the hazard (defined as the likelihood of being exposed

to hazards and as the susceptibility of an element at risk to suffer losses and damages);

- degree of resilience i.e., ability of a system to anticipate, cope with/absorb, resist and recover from the impact of a hazard or disaster.

In this view, resilience can be seen as an aspect of vulnerability. Actually, vulnerability and resilience are two sides of the same coin, where the first one focuses more on system protection and the second one on system recovery [45]. The second interpretation associates vulnerability to critical components, i.e., those whose failure causes large negative effects as the system. It can also be a flaw or weakness (inherent characteristic, including resilience) in the design, implementation, operation and/or management of a system, or its elements, that renders it susceptible to destruction or incapacitation when exposed to a hazard or threat, or reduces its capacity to resume new stable conditions [139].
The assessment of the vulnerability of a system requires an evaluation of the exposure to different kinds of hazards [141]. An *all-hazard approach* [127][99], encompassing a general view on the hazards targeting a given system is, thus, needed. In this thesis work, vulnerability is conceptualized as a global system property related to the system susceptibility to all hazards (intentional, random internal and natural) and to resilience. Eventually, the research work is developed with such reference to intentional hazards.

1.2 Multi-Criteria Decision Making (MCDM) framework of vulnerability

A broad spectrum of approaches have been proposed for system vulnerability assessment. In the all-hazard approach, malevolent acts, accidental and natural occurrences are all considered. Yet, they require a different analytical treatment. Random accidents, natural failures and unintentional man-made hazards are typically known and

categorized by emergency planners. Their occurrence can be typically modeled within a probabilistic framework typical of classical risk assessment approaches. Conversely, terrorism is a hazard that eludes a quantification by probability theory due to the intentional and malevolent planning it implies [140]: in such cases, alternative methods should be sought. The occurrence of malevolent acts brings issues related to the time frame of the analysis and to uncertainty due to behaviors of different rationality. As a result, classical risk analysis approaches can be very difficult to apply. Multi-Criteria Decision Aid (MCDA) can provide a formal procedure for the assessment. Indeed, significant advances in MCDA over the last three decades constitute a powerful non-parametric alternative methodological approach for ranking, prioritization and classification problems, which can be adopted also for the vulnerability assessment of complex systems [13].

In all generality, MCDA aims at *constructing* a systematic view of the decision maker preferences consistent with a certain set of assumptions, so as to give coherent guidance to the decision maker in the search for the preferred solution (for example, in the context of interest to the present thesis, the less/more vulnerable system configuration). This is achieved by constructing a model to represent the decision maker preferences and value judgements. The model contains two primary elements, viz.:

1. *Preferences in terms of each individual criterion,* i.e., models describing the relative importance or desirability of achieving different levels of performance for each identified criterion.

2. *Aggregation scheme*, for allowing inter-criteria comparisons, in order to combine preferences across criteria.

Then, according to the nature of the decision making problem the policy of the decision maker and the overall objective of the decision, four different analyses can be performed to provide (Figure 1-1):

1. identify the best alternative or select a limited set of the best alternatives;

2. rank-order the alternatives from best to worst ones;

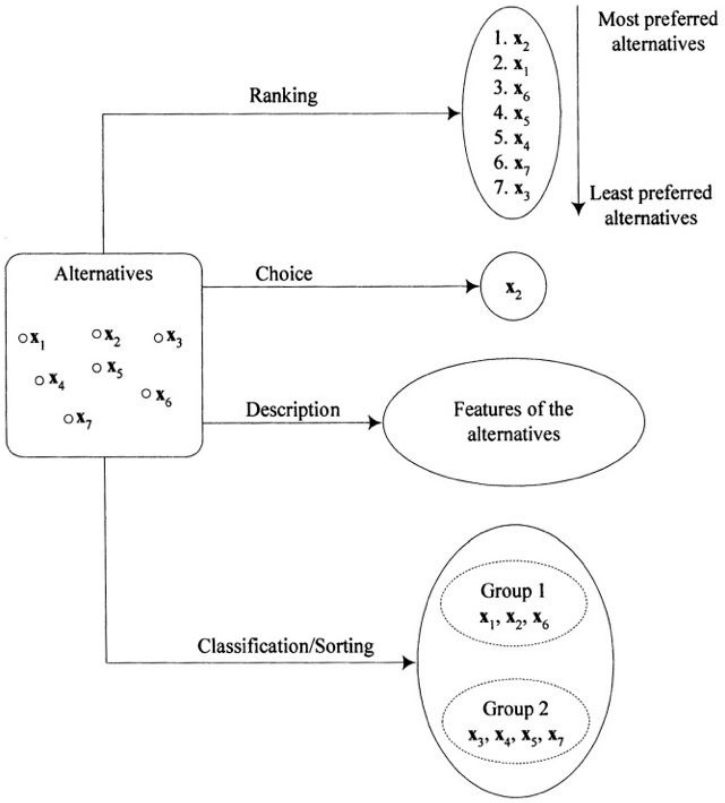

Figure 1-1: Decision making analysis

3. classify/sort the alternatives into pre-defined homogenous groups;

4. identify distinguishing features of the alternatives and perform their description based on these features.

Three different classes of preference models are here briefly presented, viz. value measurement, satisficing and aspiration-based methods and outranking.

1.2.1 Value measurement theory

The objective is to associate a real number to each alternative, in order to produce a preference order on the alternatives which is consistent with the decision maker value judgments [13]. In other words, one seeks to associate a number (or *"value"*) $V(a)$ to each alternative a, in such a way that a is judged to be preferred to b ($a \succ b$), taking all criteria into account, if and only if $V(a) > V(b)$, which also implies indifference

between a and b ($a \sim b$) if and only if $V(a) = V(b)$.

Utility theory can be viewed as an extension of value measurement theory, relating to the use of probabilities and expectations to deal with uncertainty. Multiattribute value (or utility) theory (MAVT or MAUT) is one of the more widely applied multricriteria decision analysis methods [65][68]. Ever since its origins in the late 1960's, concerns for the practical application of multi-attribute value theory (MAUT) or, more generally, multi-attibute utility theory (MAUT), have influenced developments in the field [100][38][124][120][71]. The field has benefited from the longstanding interests of psychologists, engineers, management scientists and mathematicians which has brought a countinuing awareness of behavioural and social issues. In recent years these issues have become more widely embraced by the MCDA community as a whole, as dicussed by Bouyssou et al. [19] and by Korhonen and Wallenius [74]. The problems of preference and aggregation are considered attentively. A specific method, namely ACUTA (Analytic Centre UTilité Additive) [17] is described in Chapter 2 and has been applied in the present work for ranking safety-critical systems according to their vulnerability.

1.2.2 Satisficing and aspiration-based methods

In some instances, the definition of a criterion may imply an objective ordering of the set of alternatives in terms of this criterion. Such a well-defined measure of performance is presented by a partial preference function, represented by $z_i(a)$, to show the performance level or attribute value of the alternative a according to criterion i [13]. The satisficing and aspiration-based methods operate directly on the partial preference functions without further transformation. The assumption, however, is that the preference function values have cardinal meaning, i.e., are more than simply ordinal or categorical, and relate to operationally meaningful and measurable attributes. Decision makers focus initially on seeking improvements to what is perceived to be the most important criterion. In effect, available alternatives are systematically eliminated until, in the view of the decision maker, a satisfactory level of performance for this criterion has been ensured. At this point, attention shifts to the next most

important criterion, and the search continues amongst the remaining alternatives for those which ensure satisfactory performance on this criterion. Goal programming and its variants [28][103][112][80][52][53][81] are linked to this concept.

1.2.3 Outranking

As with the satisficing models, outranking models are applied directly to partial preference functions, which are assumed to have been defined for each criterion. These preference functions may correspond to natural attributes on a cardinal scale, or may be constructed in some way, typically as oridnal or ordered categorical scales. The general principle can be seen as follows: we noted that if for two alternatives a and b, $z_i(a) \geq z_i(b)$ for all criteria i (with strict inequality $z_i(a) > z_i(b)$ for at least one criterion), then we can immediately conclude that a should be preferred to b (provided, of course, that the set of criteria is sufficiently complete). In this event, we could say that the evidence favouring the conclusion that alternative a is as good or better than alternative b is unarguable, and a is said to *dominate* b. More generally, we shall say that a *outranks* alternative b if there is "sufficient" evidence to justify a conclusion that a is at least as good as b, taking all criteria into account [106][122][123]. The two most prominent outranking approaches, the ELECTRE (ELimination Et Choix Traduisant la REalité) family of methods [107], developed by Roy and associates at LAMSADE (Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision), University of Paris Dauphine, and PROMETHEE (Preference Ranking Organization Method for Enrichment and Evaluation) [20][21], proposed by Brans from the Free University of Brussels. A simplified version of ELECTRE III method is applied in our study (Chapter 3) to assign vulnerability classes to a set of safety-critical systems (specifically, NPPs).

A number of examples of applications of MCDA approaches to the assessment/ranking/ prioritization of the vulnerability of safety-critical systems exist. Apostolakis and Lemon [5] and Patterson and Apostolakis [95] focus on the identification of critical locations in infrastructures; these are seen as geographical points that are exposed

to intentional attacks. Critical locations are not limited to individual infrastructures but may affect multiple infrastructures. For example, water and electrical distribution systems may occupy the same service tunnels. The vulnerabilities and their ranking according to potential impacts are obtained by Multi-Attribute Utility Theory (MAUT) [90].

Konce et al. [73] have proposed a methodology for ranking components of a bulk power system with respect to its risk significance to the involved stakeholders; the likelihood and the extent of power outages when components fail to perform their designed functions are analyzed; the consequences associated with the failures are determined by considering the type and number of customers affected.

Johansson and Hassel [61] have proposed a framework for considering structural and functional properties of interdependent systems and developed a predictive model in a vulnerability analysis context.

Piwowar et al. [98] have proposed a systemic analysis which accounts for malevolence, i.e., the willingness to cause damage.

Cailloux and Mousseau [25] have proposed a framework to evaluate and compare the threats and vulnerabilities associated with territorial zones according to multiple criteria (industrial activity, population, etc.) by using an adapted ELECTRE method.

1.3 Research Issues and Motivation

As mentioned above, the vulnerability of complex safety-critical systems and infrastructures (e.g., nuclear power plants) is of great concern, given the multiple and diverse hazards that they are exposed to (intentional, random and natural).

The susceptibility associated with random internal hazards and natural hazards is classically treated within a probabilistic framework to handle both the aleatory uncertainty in the occurrence of the accident events and their consequences [76] and the epistemic uncertainty on the hypotheses and parameters of the models used. Intentional hazards relate to malevolent acts and lack of a well-established methodology for accounting for uncertainty due to behaviors of different rationality [141][32]. Due to

the nature of this kind of hazard (low probability of occurence but important effects), an increased attention in building a proper framework for vulnerability analyses to guide designers, managers, and stakeholders in the decision making process is motivated.

The intentional hazards are ranged to the intended "error" of the spectrum of human-related threats, targeted malicious attacks, either physical (e.g., explosive devices) or cyber. This calls for very sophisticated models capable of describing a complex human behavior under abnormal conditions, together with systems response models to predict a potential damage to the system analyzed. There is a clear need to take increased economic pressure into account when assessing human response behavior, to modify maintenance strategies and spare-parts managing, to focus on the search for entry points for hacker attacks after the industrial control systems have become more open and less dedicated, etc [76].

With respect to all that mentionned above, the analysis is difficult to perform by classical risk assessment methods [76][7][9]. For this reason, a combination of two disciplines, namely risk analysis and multi-criteria decision-making, is strongly advised. The contribution of the present Ph.D. in this framework of analysis is detailed in the next Section.

1.4 Synthesis of the contribution of the Thesis

In this Ph.D. thesis, the quantitative assessment and management of the vulnerability to intentional hazards of complex systems are considered. This includes: (i) the *representation* and *modeling* of the susceptibility to malevolent acts within a hierarchical framework; (ii) the quantitative *assessment* of the vulnerability; (iii) the optimal choice of the protective actions to reduce the vulnerability level (*management*). These tasks are performed within a framework that combines the disciplines of risk analysis and MCDA. Applications are given with reference to the nuclear power plants.

The following contributions have emerged from the work performed during this Ph.D.:

1. a hierarchical model is developed to characterize the susceptibility of safety-critical systems to intentional hazards. This allows the systematic identification of the main sources of vulnerability for the systems under analysis.

2. vulnerability is evaluated using the hierarchical framework described above, in two ways:

    a. the vulnerabilities of a group of safety-critical systems are composed and ranked by an empirical method based on the Analytic Center UTilité Additive (ACUTA) approach [17].

    b. since the ACUTA approach can only provide a ranking (relative evaluation) of the vulnerability, an empirical classification model, based on the MR-Sort (Majority Rule sorting) method [82] is used to provide a quantitative (absolute) assessment of the vulnerability of each system of interest. The MR-Sort classification model contains a group of (adjustable) parameters that have to be calibrated by means of a set of *empirical* classification examples (also called training set), i.e., a set of alternatives with the corresponding pre-assigned vulnerability classes.

3. due to the finite (typically small) size of the set of training classification examples usually available in the analysis of real complex safety-critical systems, the performance of the classification model can be impaired. A quantitative assessment of the performance of the classification model (in terms of accuracy and confidence in the assignments) is needed, to account for the uncertainty introduced into the analysis by the empirical construction of the vulnerability classification model. This has been analyzed by three different approches, namely, the model-retrieval-based method [82], the Bootstrap method [35] and the leave-one-out cross-validation technique [11].

4. the classification examples provided by the experts for the construction of the classification model may contain contradictions: a validation of the consistency of the data set is, thus, opportune. In this thesis, two approaches are used to

tackle this issue: in particular, the inconcsistencies in the data examples are "resolved" by deleting or relaxing, respectively, some constraints in the process of model construction [82].

5. a set of protective actions should be chosen to effectively reduce the level of vulnerability of the critical system under consideration: two frameworks, based on sensitivity indicators and optimization-based approaches are proposed to this aim. In particular, sensitivity indicators are originally introduced as measures of the variation in the vulnerability class that a safety-critical system is expected to undergo after the application of a given set of protective actions. These indicators form the basis of an algorithm to rank different combinations of actions according to their effectiveness in reducing the safety-critical systems vulnerability. Then, three different optimization approaches have been explored: (i) one single classification model is built to evaluate and minimize system vulnerability; (ii) an ensemble of compatible classification models, generated by the bootstrap method, is employed to perform a "robust" optimization, taking as reference the "worst-case" scenario over the group of models; (iii) finally, a distribution of classification models, still obtained by bootstrap, is considered to address vulnerability reduction in a "probabilistic" fashion (i.e., by minimizing the "expected" vulnerability of a fleet of systems).

In Chapters 2-5, all the above objectives are discussed, together with their relevance to each case study described in the papers (cf Part II).

1.5 Structure of the Thesis

The thesis is composed of two parts: Part I, subdivided in six Chapters, introduces the current issues and challenges pertinent to vulnerability analysis of complex safety-critical systems, describes the research objectives undertaken, illustrates the methods developed and applied in this Ph.D. work, discusses some of the results obtained in the case studies and provides general conclusions and some future work perspectives.

Part II is a collection of five selected papers, scientifically reporting on the outcomes of the research work performed during the thesis, which the readers are referred to for further details. Table 1.1 summarizes the thesis structure.

Chapter 2 starts with a brief critical discussion of the approaches based on decision making ranking theory, that have been employed for the analysis of safety-critical systems. Then, the hierarchical modelling framework for representing and describing the vulnerability of complex energy systems to intentional hazards is proposed, which can be leveraged efficiently to facilitate the management of complexity in the analysis of large-scale complex systems. The empirical method based on ACUTA, used to compare and rank the vulnerability of the complex systems is presented.

In Chapter 3, methods to provide an absolute quantitative evaluation of the vulnerability of the complex systems are considered. A study of available approaches is presented. Then, the problem is formulated based on an empirical classification framework using the majority rule sorting method (MR-Sort), a quantitative evaluation of the vulnerability level of different critical systems is given. In addition, three different methods are used to assess the performance of this classification model with respect to its accuracy and confidence in the assignments (i.e., a model-retrieval-based method, the Bootstrap method and the leave-one-out cross-validation technique). The methods used to deal with inconsistent data are also presented.

Chapter 4 focuses on the inverse multicriteria classification problem, i.e., the selection of a set of actions able to reduce the vulnerability (class) of a group of systems, under given constraints. The problem is tackled by two perspectives, the first based on a novel sensitivity indicator and the second on optimization. In particular, for the second perspective, an optimization framework is proposed for properly selecting protective actions in order to minimize the overall vulnerability of a group of safety-critical systems.

Chapter 5 contains the applications of the proposed models and methodologies to fictitious and real NPPs. Chapter 6 draws the conclusions of this Ph.D. study and presents relevant open issues and perspectives for future research. Figure 1-2 provides a pictorial view of the issues and the approaches considered in the present Ph.D. work

on vulnerability analysis of safety-critical systems.

Part II of this thesis includes the collection of published and submitted papers, which constitute the pillars of the present doctoral thesis. Paper (i) presents the hierarchical representation framework and its application to vulnerability analysis based on decision making ranking theory (see Chapter 2 of Part I). Papers (ii) and (iii) concern the classification of the vulnerability level of NPPs (see Chapter 3 and Chapter 5). Specifically, Paper(ii) addresses this analysis by also providing the accuracy and confidence of the classification assignments. Paper (iii) gives a solution of how to deal with the inconsistencies that may exist in the data available to build the classification model. Papers (iv) and (v) form the basis for the study of inverse classification problem detailed in Chapter 4. The introduction of a sensitivity indicator, used as measures of the variation in the vulnerability class that a safety-critical system is expected to undergo after the application of a given set of protective actions, is the main contributions of Paper (iv). Finally, for Paper (v), the issue of selecting protective actions to reduce the vulnerability of safety-critical systems has been tackled within different optimization-based frameworks relying on the empirical classification model presented in Paper (ii).

Table 1.1: Structure of the thesis

Topic	**PART I Chapter(s)**	**PART II Paper(s)**
Vulnerability of critical systems	**2**	**i**
Hierarchical modeling		i
Decision making ranking theory		i
·Analytic Centre UTilité Additive method(ACUTA)		
Vulnerability assessment	**3**	**ii-iii**
Decision making sorting theory		ii
·Majority Rule sorting classification method(MR-Sort)		
Accuracy and confidence of the classification model		ii
·model-retrieval-based method		
·bootstrap method		
·leave-one-out cross-validation technique		
Study of inconsistency in data		iii
·constraints deletion approach		
·constraints relaxation approach		
Vulnerability management– Inverse classification problem	**4**	**iv-v**
Sensitivity indicators		iv
Optimization-based approaches		v
·simple optimization		
·robust optimization		
·probabilistic optimization		
Applications	**5**	**i-v**
Susceptibility to intentional hazards		i,ii,iv,v
Overall level of safety-related criticality		iii

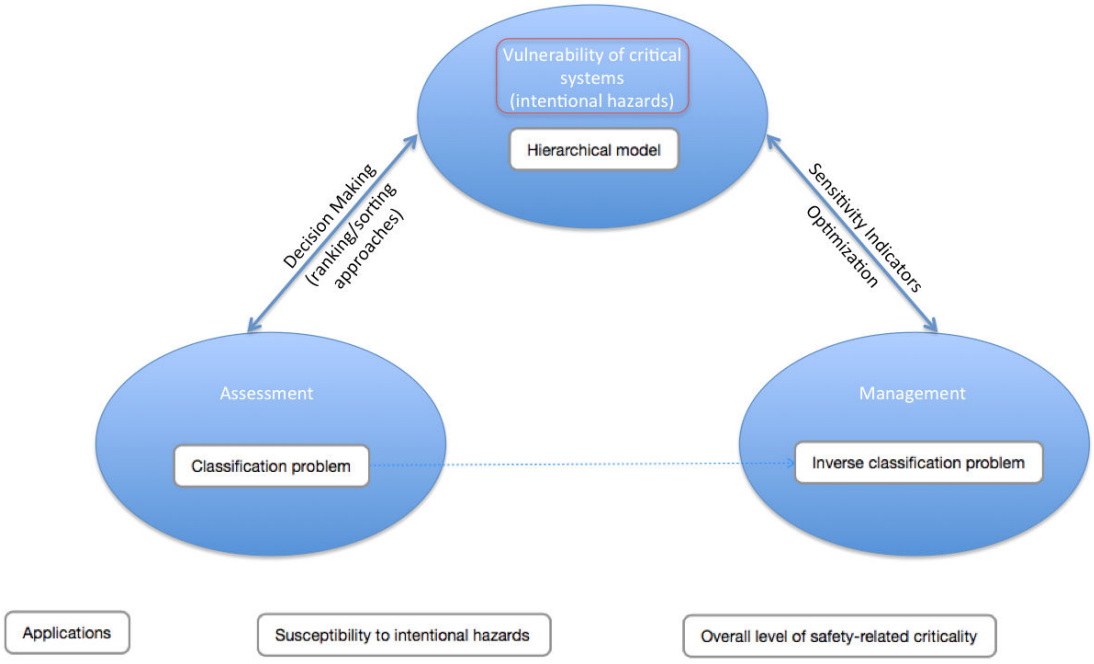

Figure 1-2: Pictorial view of the flow (topic; focus, applications and outputs) of the present Ph.D. work on vulnerability analysis of safety-critical systems.

Chapter 2

HIERARCHICAL DECISION MAKING FRAMEWORK FOR RANKING THE VULNERABILITY TO INTENTIONAL HAZARDS

"A problem well structured is a problem half solved" is an oft-quoted statement which is highly pertinent to the use of any form of modelling. Before any analysis can begin, the various stakeholders, including facilitators and technical analysts, need to develop a common understanding of the problem, of the decisions that have to be made, and of the criteria by which such decisions are to be judged and evaluated.

The modelling of any real-life system for vulnerability analysis proposes well-defined system criteria and usually simplifications of the system representation determined by the context in which the model is used. The aim of this chapter is to critically review previous inspiring research regarding the modelling of complex systems, as well as to describe the author's proposed modelling approach based on a decision making ranking theory. In particular, the first Section briefly introduces the existing system representation techniques and the decision making ranking theory. The second Section details the general hierarchical modelling framework used in this thesis for

representing the vulnerability of a complex energy system.

2.1 State of the art

2.1.1 Overview on the existing system representation techniques

Several types of system representation approaches exist in literature and they rely mainly on a hierarchy or graph structure.

Hierarchical Modelling have been often adopted to represent and model complex systems, since many organizational and technology-based systems are hierarchical in nature [46]. "Frequently, complexity takes form of hierarchy, whereby a complex system is composed of interrelated subsystems that have in turn their own subsystems, and so on, until some lowest level of elementary components is reached" [31]. This approach can be based on different perspectives, e.g., functional, technical, organizational, geographical, political, etc., and can allow simplifying the modelling process and the ultimate management of the system as a whole [46].

Hierarchical functional models include Goal Tree Success Tree (GTST) – also combined with Master Logic Diagram (MLD) – and Multilevel Flow Modelling (MFM). The GTST is a functional hierarchy of a system organized in levels starting with a goal at the top; the MLD, developed and displayed hierarchically, shows the relationships among independent parts of the systems: the combined GTST – MLD provides a powerful functional/structural description method. Finally, the dynamic version of the approach, namely the GTST – Dynamic MLD (GTST-DMLD), allows describing the temporal behavior of the systems [108]. Multilevel Flow Models [84][85], developed in the field of artificial intelligence, have been proposed for qualitative reasoning, i.e., for representing and structuring knowledge about physical phenomena and systems. They consider cause-effect relations and facilitate the reasoning at different levels of abstraction on the basis of "means-end" and "whole-part" decomposition and aggregation procedures. Goals, functions and flow of material, energy and information are

connected to form a hyper graph. They are mainly used for measurement validation (e.g., for checking the measurement of a mass or energy flow), alarm analysis (e.g, for the identification of primary and secondary alarm), and fault diagnosis (i.e., for the identification of the consequences and the root causes of a disturb in the system functioning) [78].

In risk analysis, common representation techniques are hierarchical trees that are commonly used to identify (i) the initiating causes of a pre-specified, undesired event or (ii) the accident sequences that can generate from a single initiating event, through the development of structured logic trees, i.e., fault and event trees, respectively [138].

In complex network theory, instead, complex systems are represented by networks where the nodes stand for the components and the links describe the physical and relational connections among them. Network-based approaches model interdependent critical infrastructures (CIs) on the basis of their topologies or flow patterns by topology-based and flow-based methods, respectively.

Probabilistic modelling includes Petri nets, Bayesian networks and flowgraphs. A Petri net is a directed graph that consists of places (i.e., conditions), transitions (i.e., events that may occur) and directed arcs describing which places are pre- and/or post-conditions for each transition. They are well suited for modelling the behavior of distributed systems. Laprie et al. have adopted Petri nets to describe and analyze high level scenarios that may take place when failures occur in two interdependent infrastructures (i.e., in information and electricity infrastructures), considering the effect that failures in one infrastructure have on the other and accounting also for malicious attacks [77]. Bayesian networks are based on directed acyclic graphs where nodes are random variables representing the state of components and edges are conditional dependencies, reflecting the causal relationships among adverse events. Classical Bayesian networks provide static models of the system at each time step; however, recently, dynamic Bayesian networks have been introduced [94]. Differently, flowgraphs model the outcomes of random variables: in this framework, nodes identify the actual physical state of a system and edges model the allowable transitions, the probabilities of different outcomes, and waiting times until the occurrence of out-

comes of interest [51].

2.1.2 Decision making ranking theory

At this stage, let us assume that the problem structuring phases have generated a set of alternatives (Which may be a discrete list of alternatives, or may be defined implicitly by a set of constraints on a vector of decision variables), and a set of criteria against which these alternatives are to be evaluated and compared (also based on the chosen representation technique).

In the chapter, we consider the problem of vulnerability analysis as a "ranking problem". In this context, the UTA family of methods is one of the most important one (see also Jacquet-Lagrèze and Siskos [59]; Belton and Stewart [13]; Bouyssou et al. [18] and Siskos et al. [114], for a general overview and Jacquet-Lagrèze [58]; Siskos et al. [115], for software implementations). UTA methods are value-focused approaches, i.e., they are based on the theory of multiattribute value functions or MAVT (Keeney and Raiffa [68]), which rests on the hypothesis that any person facing a multiple criteria decision problem intuitively attempts to maximize a function that aggregates all criteria into a global evaluation of each alternative. MAVT-based methods differ both in the way criteria are modeled and in the way values are aggregated; UTA methods, in particular, model the perceived value of each criterion, called marginal value function, by using piecewise linear functions and use the additive model in order to aggregate them into a global value function.

2.2 Hierarchical framework for vulnerability analysis

2.2.1 Proposed framework

Vulnerability is defined in different ways depending on the domains of application, e.g.: vulnerability is a measure of possible future harm due to exposure to a haz-

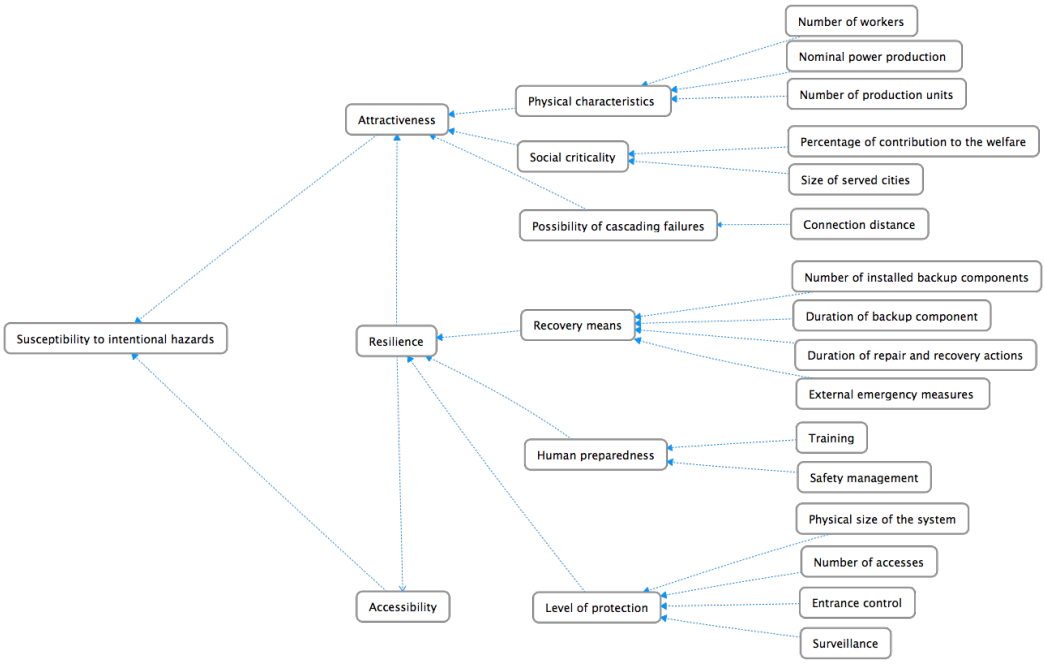

Figure 2-1: Hierarchical model Susceptibility to intentional hazards

ard [76]; the identification of weaknesses in security, focusing on defined threats that could compromise a system ability to provide a service [93]; the set of conditions and processes resulting from physical, social, economic, and environmental factors, which increase the susceptibility of a community to the impact of hazards [47]. With the focus on the susceptibility to intentional hazards, a four-layers hierarchical model is built as shown in Figure 2-1. The idea behind the use of hierarchical representation is that most complex systems are in the form of a hierarchy (see the previous Section 2.1). The Hierarchical Modelling allows studying the system at different level of details and extracting from each level the groups of criteria that are critical from the safety viewpoint. The evaluation through the framework is shown by way of analysing the susceptibility to intentional hazards of a safety-critical system, namely a Nuclear Power Plant (NPP), considering the vulnerability sources and the related features, the system technical and physical features, and the dependencies and interdependencies on other systems.

The susceptibility to intentional hazards is characterised in terms of attractiveness

and accessibility. These are hierarchically broken down into factors which influence them, including resilience seen as pre-attack protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness). The decomposition is made in 6 criteria ($MCrit = \{MCrit_1, MCrit_2, ..., MCrit_i, ..., MCrit_n\}$ with $n = 6$: physical characteristics ($MCrit_1$), social criticality ($MCrit_2$), possibility of cascading failures ($MCrit_3$), recovery means ($MCrit_4$), human preparedness ($MCrit_5$) and level of protection ($MCrit_6$)) which are further decomposed into a layer of basic subcriteria, for which data and information can be collected to make their evaluation. In what follows, a description of the second layer criteria (attractiveness and accessibility) is given.

Attractiveness

This second-layer criterion is intended to capture the interest that terrorists may have to attack the system. Such interest is considered to be driven mainly by the effects that the attack can cause, which include damages to the assets and environment, injured people, deaths. These depend on the physical characteristics of the system, its social criticality, the possibility of cascading effects and the system resilience. In a general sense, resilience represents the ability to avoid the occurrence of accidents despite the persistence of poor circumstances or to recover from some unexpected events [41]. It is the ability of a system to anticipate, cope with/absorb, resist and recover from the impact of a hazard (technical) or a disaster (social). Resilience reflects a dynamic confluence of factors that promotes positive adaptation despite exposure to adverse life experiences. In our model, it is presented in terms of capacity of recovery, human preparedness and level of protection.
The preference direction characterising this factor is such that the more attractive the system is, the more it should be protected.

Accessibility

Accessibility is introduced as a criterion in the second layer of the hierarchy to describe the degree to which it is easy or difficult to arrive at a system in order to intentionally

damage it. It is a function of resilience through the level of protection present to defend against malevolent attacks.
Each third-layer criterion is constituted by several subcriteria (Table 2.1). The value of the subcriteria can be crisp numbers or language terms according to the contents. Each of the subcriterion is analysed in giving an explanation of the contribution on the corresponding third-layer criterion. More details of the third-layer criteira can be found in appended Papers (i) and (ii).

2.2.2 Decision making methodology for ranking the susceptibility to intentional hazards

The hierarchical model just presented structures the susceptibility of a critical system to intentional attacks in terms of a number of criteria. The 16 basic, bottom-layer subcriteria are organised into 6 main ones: the physical characteristics, the social criticality, the possibility of cascading failures, the recovery means, the human preparedness and the level of protection.
For the quantitative assessment, each of the 16 basic subcriteria needs to be assigned a value function in relation to the main criterion to which it contributes. The criteria of the layers are defined and assigned preference directions for treatment in the decision-making process (Table 2.1). The preference direction of a criterion indicates towards which state it is desirable to lead it to reduce susceptibility, i.e., it is assigned from the point of view of the defender of an attack who is concerned with protecting the system. Although only the 6 criteria in the third level of the hierarchy will be considered in the exemplary demonstration on the NPPs evaluation, examples of scales of evaluation also of the basic subcriteria of the last layer are proposed, in relation to the characteristics of NPPs for exemplification purposes. The assignment can be done in relative terms, by comparing different systems with different characteristics. We perform a decision-making process for the evaluation of their characteristics with respect to susceptibility to intentional attacks. Special focus is given to the preference disaggregation analysis (PDA) of MCDA. The preference disaggregation

Table 2.1: Criteria, subcriteria and preference directions

Criterion	**Physical characteristics**	**Social criticality**	**Possibility of cascading failures**
Subcriteria	Number of workers	Percentage of contribution to the welfare	Connection distance
	Nominal power production	Size of served cities	
	Number of production units		
Preference direction	Min	Min	Min

Criterion	**Recovery means**	**Human preparedness**	**Level of protection**
Subcriteria	Number of installed backup components	Training	Physical size of the system
	Duration of backup components	Safety management	Number of accesses
	Duration of repair and recovery actions		Entrance control
	External emergency measures		Surveillance
Preference direction	Max	Max	Max

approach refers to the analysis (disaggregation) of the global preferences (judgement policy) of the decision maker in order to identify the criteria aggregation model that underlies the preference result. It can be used to analyze the actual decisions taken by the decision maker so that an appropriate model can be constructed representing the decision maker system of preferences, as consistently as possible. As so, we first build a ranking of fictitious NPPs, through the authors' subjective preferential judgment of indirect data. This serves for constructing the basis for the relative evaluation of the characteristics of real NPPs. To carry out the decision-making process for the evaluation, we resort to a multiple criteria decision aid (MCDA) technique named ACUTA (Analytic Centre UTilité Additive) based on the computation of the analytic centre of a polyhedron for the selection of additive value functions that are compatible with holistic assessments of the preferences in the criteria [17]. Being central by definition and uniquely defined, the analytic centre benefits from theoretical advantages over the notion of centrality used in other meta-UTA methods. A brief explanation of the method is given as follow.

Analytic Center

The idea of the analytic centre of a polyhedron was first introduced by Huard [49] and later reintroduced by Sonnevend [117] in the context of convex optimization techniques. The theoretical framework around this concept lies at the heart of interior-point methods for solving linear programming optimisation problems. In ACUTA, it

is suggested to compute a unique, well-defined and central solution for aggregation-disaggregation methods based on additive piecewise linear value function models [17].

ACUTA

The UTA(UTilité Additive) method consists in building a piecewise linear additive decision model from a preference structure using linear programming. Let x be the set of possible alternatives and $x_L = \{x_p, p = 1, ..., N\}$ the learning set. In x_L, alternatives are ranked in order of decreasing preference by the DM (Decision Maker), i.e., $x_p \succsim x_{p+1}, p = 1, ..., N-1$, where $\succsim$ expresses that x_p is either preferred ($\succ$) or indifferent ($\sim$) to x_{p+1}. The values of the n criteria, denoted by $MCrit_i (i = 1, ..., n)$, belong to the interval $[\overline{\chi_i}, \underline{\chi_i}]$ that, for each i, corresponds to the range between the worst ($\underline{\chi_i}$) and best ($\overline{\chi_i}$) values found for attribute i among the alternatives in x. Our purpose is to establish marginal value functions $\nu_i(\chi_i)$ for each criterion in order to model the perceived value of each alternative. Since these values are piecewise linear functions, the range of values on each criterion is divided into subintervals using a predefined number of α_i points such that $\chi_i = \{\underline{\chi_i} = \chi_i^1, \chi_i^2, ..., \chi_i^{\alpha_i} = \overline{\chi_i}\}$. The subdivision makes it possible to compute value functions by linear interpolation between the values $\nu_i(\chi_i^l)$ that have to be estimated and hence appear as variables in the linear program. Using the degrees of freedom in the definition of a value function, we set $\nu_i(\underline{\chi_i}) = 0$ and

$$\sum_{i=1}^{n} \nu_i(\overline{\chi_i}) = 1 \tag{2.1}$$

This implies that $\nu_i(\overline{\chi_i})$ can be interpreted as the tradeoff associated to criterion i. Furthermore, all value functions should be monotonic, that is $\nu_i(\chi_i^{l+1}) - \nu_i(\chi_i^l) \geqslant \lambda (\forall\ i\ and\ l = 1, ..., \alpha_i - 1)$, with $\lambda \geqslant 0$. According to the additive model, the global value $\nu(x_p)$ of an alternative x_p is given by the sum of its marginal values. In other terms, if the value of the p^{th} alternative on attribute i is denoted by x_{ip} , the global

value of x_p is given by

$$\nu(x_p) = \sum_{i=1}^{n} \nu_i(x_{ip}) \tag{2.2}$$

This analytic expression of an alternative's global value allows for modelling the preferences of the DM, as expressed in the ranking of the learning set, using the following linear constraints, which we call preference constraints:

$$\nu(x_p) - \nu(x_{p+1}) \geqslant \delta \qquad if\ x_p \succ x_{p+1}, \tag{2.3}$$

$$\nu(x_p) - \nu(x_{p+1}) = 0 \qquad if\ x_p \sim x_{p+1}. \tag{2.4}$$

Here, λ is a positive number, called preference threshold, which is usually set to a small value. The assessment of the $\nu_i(\chi_i^l)$ variables should be done in such a way that the deviation from the preferences expressed by the DM in the subset x_L is minimal. The adaptation of the linear additive aggregation-disaggregation model to the analytic centre formulation is quite straightforward and gives rise to the ACUTA method; the introduction of slack variables into the objective function leads to the following nonlinear optimisation problem, which can be solved without further modifications:

$$max \sum_{p=1}^{N-1} ln(s_j) = \sum_{i=1}^{n} \sum_{l=1}^{x_i - 1} ln(s_{il}), \tag{2.5}$$

$$s.t. \quad \nu(x_p) - \nu(x_{p+1}) = 0 \quad if\ x_p \sim x_{p+1}, \tag{2.6}$$

$$(\nu(x_p) - \nu(x_p + 1)) - \delta = s_p \quad if\ x_p \succ x_{p+1}, \tag{2.7}$$

$$s_{il} = (\nu(\chi_i^{l+1}) - \nu(\chi_i^l)) - \lambda, \tag{2.8}$$

$$\sum_{i=1}^{n} \nu_i(\overline{\chi_i}) = 1. \tag{2.9}$$

Since this approach maximises the sum of slacks, parameters δ and λ can be omitted, and this is considered an advantage. The essential advantage of this method, however, is the centrality and uniqueness of the solutions it produces.

Chapter 3

CLASSIFICATION MODEL FOR THE QUANTITATIVE ASSESSMENT OF THE VULNERABILITY TO INTENTIONAL HAZARDS

As mentioned previously, due to the specific features (low frequency but important effects) of intentional hazards (characterised by significant *uncertainties* due to behaviours of different rationality) the analysis is difficult to perform by traditional risk assessment methods [76][7][9]. For this reason, in the present thesis work we propose to tackle the issue of evaluating vulnerability to malevolent intentional acts by an empirical classification modelling framework. Classification refers to the assignment of a finite set of alternatives into predefined groups. It can provide many benefits to an organization such as reducing decision making time, improving the consistency of decisions, and reducing dependence on scarce human experts. In our case, it can give an absolute judgement of the vulnerability to intentional hazards of each single alternative rather than a relative one (e.g., given by a rank).

For several decades multivariate statistical analysis techniques such as discriminant analysis (linear [39] and quadratic [116]), and econometric techniques such as logit and probit analysis [87][88], the linear probability model, etc., have dominated the field of classification problem. However, the parametric nature and the statistical assumptions/restrictions of such approaches have been an issue of major criticism and skepticism on the applicability and the usefulness of such methods in practice.

The continuous advances in other fields including operations research and artificial intelligence led many scientists and researchers to exploit the new capabilities of these fields, in developing more efficient classification techniques. Among the attempts made one can mention neural networks [118][96][50][75][6], machine learning [72][62][43], fuzzy sets [131][56][57][54][12] as well as multicriteria decision aid. Multicriteria decision aid (MCDA) has several distinctive and attractive features, involving, mainly, its decision support orientation. The significant advances in MCDA over the last three decades constitute a powerful non-parametric alternative methodological approach to study classification problems [33].

3.1 State of the art

The focus of this study is based on Multicriteria decision aid (MCDA), which is an advanced field of operations research and has evolved rapidly over the past three decades both at the research and practical level.

The development of the MCDA field has been motivated by the simple finding that resolving complex real-world decision problems cannot be performed on the basis of unidimensional approaches. However, when employing a more realistic approach considering all factors relevant to a decision making situation, one is faced with the problem referring to the aggregation of the existing multiple factors. The complexity of this problem often prohibits decision makers from employing this attractive approach.

The ELECTRE family of methods (ELimination Et Choix Traduisant la REalité [104]) developed by Bernard Roy during the late 1960s set also the foundations of

the outranking relation theory (ORT). Since then, it has been widely used by MCDA researchers, mainly in Europe. The family of ELECTRE methods differ according to the degree of complexity, or richness of the information required or according to the nature of the underlying problem or problematique. The ELECTRE TRI method [130] is a member of this family of methods, developed for addressing classification problems. A simplified version of this method is applied in this work and presented in the following sections.

3.2 The majority rule sorting method (MR-Sort)

The Majority Rule Sorting (MR-Sort) method is a simplified version of ELECTRE TRI, an outranking sorting procedure in which the assignment of an alternative to a given category is determined using a complex concordance non-discordance rule [105][92].

3.2.1 The MR-Sort algorithm

We assume that the alternative to be classified (e.g., a safety-critical system or infrastructure of interests, e.g., a nuclear power plant) can be described by an n-tuple of elements $x = \{x_1, x_2, ..., x_i, ..., x_n\}$, which represent the evaluation of the alternative with respect to a set of n criteria (by way of example, in the present work the criteria used to evaluate the vulnerability of a safety critical system of interest may include its physical characteristics, social criticality, level of protection and so on: as presented by the hierarchical model in the previous Chapter). We denote the set of criteria by $MCrit = \{MCrit_i, i \in \{1, 2, ..., n\}\}$ and assume that the values x_i of criterion i range in the set X_i [102] (e.g., all the criteria range in [0, 1]). The MR-Sort procedure allows assigning any alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\} \in X = X_1 \times X_2 \times ... \times X_i \times ... \times X_n$ to a particular pre-defined category (in this work, a class of vulnerability), in a given ordered set of categories, $\{A^h : h = 1, 2, ..., M\}$; $M = 4$ categories are considered in this work: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, A^4 = serious.
To this aim, the model is further specialised in the following way:

- We assume that X_i is a subset of $\mathbb{R}$ for all $i \in \mathbb{N}$ and the sub-intervals $(X_i^1, X_i^2, ..., X_i^h, ..., X_i^M)$ of X_i are compatible with the order on the real numbers, i.e., for all $x_i^1 \in X_i^1, x_i^2 \in X_i^2, ..., x_i^h \in X_i^h, ..., x_i^M \in X_i^M$, we have $x_i^1 > x_i^2 > ... > x_i^h > ... > x_i^M$. We assume furthermore that each interval $X_i^h, h = 1, 2, ..., M-1$ has a smallest element b_i^h, which implies that $x_i^h \geq b_i^h > x_i^{h+1}$. The vector $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}$ (containing the lower bounds of the intervals X_i^h of criteria $i = 1, 2, ..., n$ in correspondence of category h) represents the lower limit profile of category A^h.

- There is a weight ω_i associated with each criterion $i = 1, 2, ..., n$, quantifying the relative importance of criterion i in the vulnerability assessment process; notice that the weights are normalised such that $\sum_{i=1}^{n} \omega_i = 1$.

  In this framework, a given alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\}$ is assigned to category $A^h, h = 2, ..., M-1$, iff

$$\sum_{i \in \mathbb{N}: x_i \geq b_i^h} \omega_i \geq \lambda \; and \sum_{i \in \mathbb{N}: x_i \geq b_i^{h-1}} \omega_i < \lambda, \tag{3.1}$$

  where λ is a threshold ($0 \leq \lambda \leq 1$) chosen by the analyst. Rule (3.1) is interpreted as follows. An alternative x belongs to category A^h if: 1) its evaluations in correspondence of the n criteria (i.e., the values $\{x_1, x_2, ..., x_i, ..., x_n\}$) are at least as good as b_i^h (lower limit of category A^h with respect to criterion $i, i = 1, 2, ..., n$), on a subset of criteria that has sufficient importance (in other words, on a subset of criteria that has a weight larger than or equal to the threshold λ chosen by the analyst); and at the same time 2) the weight of the subset of criteria on which the evaluations $\{x_1, x_2, ..., x_i, ..., x_n\}$ are at least as good as b_i^{h-1}(lower limit of the successive category A^{h-1} with respect to criterion $i, i = 1, 2, ..., n$), is not sufficient to justify the assignment of x to the successive category A^{h-1}.

  Notice that alternative x is assigned to the best category A^1 if $\sum_{i \in \mathbb{N}: x_i \geq b_i^1} \omega_i \geq \lambda$ and it is assigned to the worst category A^M if $\sum_{i \in \mathbb{N}: x_i \geq b_i^{M-1}} \omega_i < \lambda$. Finally, it is

straightforward to notice that the parameters of such a model are the $(M-1)\cdot n$ lower limit profiles (n limits for the $M-1$ categories), the n weights of the criteria $\omega_1, \omega_2, ..., \omega_i, ..., \omega_n$, and the threshold λ, for a total of $n \cdot M + 1$ parameters.

3.2.2 Constructing the MR-Sort classification model

In order to construct an MR-Sort classification model, we need to determine the set of $n \cdot M + 1$ parameters described in the previous subsection, i.e., the weights $\omega = \{\omega_1, \omega_2, ..., \omega_n\}$, the lower profiles $b = \{b^1, b^2, ..., b^h, ..., b^{M-1}\}$, with $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}, h = 1, 2, ..., M-1$, and the threshold λ; in this thesis work, λ is considered a fixed, constant value chosen by the analyst (e.g., λ=0.9).

As presented in the previous Chapter, a disaggregation process is thus considered. The decision maker provides a training set of classification examples $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N\}$, i.e., a set of N alternatives $x_p = \{x_1^p, x_2^p, ..., x_i^p, ..., x_n^p\}, p = 1, 2, ..., N$ together with the corresponding real pre-assigned categories (i.e., vulnerability classes) Γ_p^t (the superscript t indicates that Γ_p^t represents the true, a priori-known vulnerability class of alternative x_p).

The calibration of the $n \cdot M$ parameters is done through the learning process detailed in [82]. In extreme synthesis, the information contained in the training set D_{TR} is used to restrict the set of MR-Sort models compatible with such information, and to finally select one among them [82]. The a priori-known assignments generate constraints on the parameters of the MR-Sort model. In [82], such constraints have a linear formulation and are integrated into a Mixed Integer Program (MIP) that is designed to select one (optimal) set of such parameters ω^* and b^* (in other words, to select one classification model $M^*(\cdot|\omega^*, b^*)$ that is coherent with the data available and maximises a defined *objective function*. In [82], the optimal parameters ω^* and b^* are those that maximise the value of the minimal slack in the constraints generated by the given set of data D_{TR}. Once the (optimal) classification model $M^*(\cdot|\omega^*, b^*)$ is constructed, it can be used to assign a new alternative x (i.e., a new nuclear power plant) to one of the vulnerability classes $A^h, h = 1, 2, ..., M$: in other words, $M^*(x|\omega^*, b^*) = \Gamma_x^{M^*}$ where $\Gamma_x^{M^*}$ is the class assigned by model $M^*(\cdot|\omega^*, b^*)$ to

alternative x and assumes one value among $\{A^h : h = 1, 2, ..., M\}$. Further mathematical details about the training algorithm are not given here for brevity, they can be found in [82] and appended Paper (ii).

3.2.3 Dealing with inconsistency in the training data

As highlighted before, sorting models consist in assigning alternatives evaluated on several criteria to ordered categories. To implement such models, it is necessary to set the values of the preference parameters used in the model. Rather than fixing the values of these parameters directly, a usual approach is to infer these values from assignment examples provided by experts and decision makers (DMs). However, assignment examples provided by experts and DMs can be *inconsistent*, i.e., may not "produce" any meaningful classification model. Such a situation can be understood according to two perspectives: either the examples provided by the DM contradict each other, or the preference model is not flexible enough to account for the way the DM assigns alternatives holistically. In the first case, the DM would acknowledge a misjudgment and would agree to reconsider his/her examples; in the second case, the DM would not agree to change the examples and the preference model should be changed. In both cases, we refer to an inconsistency situation. In any case, the DM needs to know what causes inconsistency, i.e., which judgments should be changed if the aggregation model is to be kept (which is our case) [92][91].

The MIP algorithm summarized in the previous section may prove infeasible in case the assignments of the alternatives in the learning set are incompatible with all MR-sort models. In order to help the DMs to understand how their inputs are conflicting and to question previously expressed judgments, to learn about their preferences as the interactive process evolves, we formulate two MIPs that are able to: (i) find one MR-sort model that maximize the number of learning set alternatives correctly assigned and (ii) propose accordingly a possible modification for each of the conflicting alternatives.

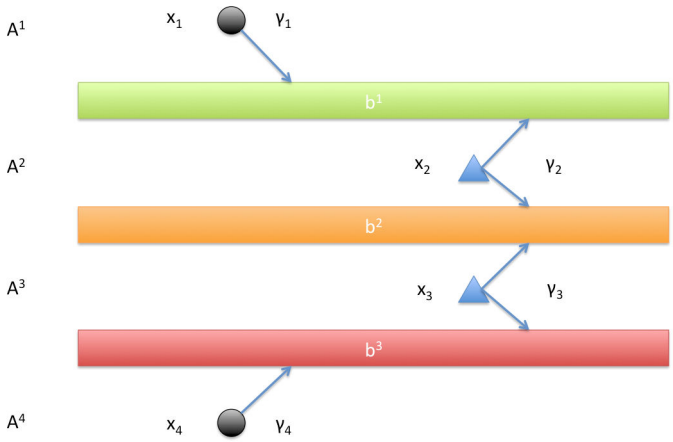

Figure 3-1: Representation of constraints deletion algorithm

Inconsistency resolution via constraints deletion

Resolving the inconsistencies can be performed by deleting a subset of constraints related to the inconsistent alternatives. As shown in Figure 3-1, we take $M = 4$, each alternative x_p can provide one or two constraints with respect to its assignment: for example, alternatives assigned to extreme categories, i.e., A^1 and A^4, provide one constraint, whereas alternatives assigned to intermediate categories, i.e., A^2 and A^3, introduce two constraints. Let us introduce a binary variable γ_p for each alternative x_p, which is equal to 1 if *all* the constraints associated to x_p are fulfilled, and equal to 0 otherwise.

The algorithm proceeds by "deleting" (i.e., removing) those constraints (i.e., those alternatives) that do not allow the creation of a compatible classification model, while maximizing the number of alternatives retained in the learning set (i.e., in minimizing the number of alternatives that are not taken into account): by so doing, we maximize the quantity of information that can be used to generate a classification model correctly. In other words, we obtain a MIP that yields a subset $D^*_{TR} \subseteq D_{TR}$ of maximal cardinality that can be represented by an MR-sort model.

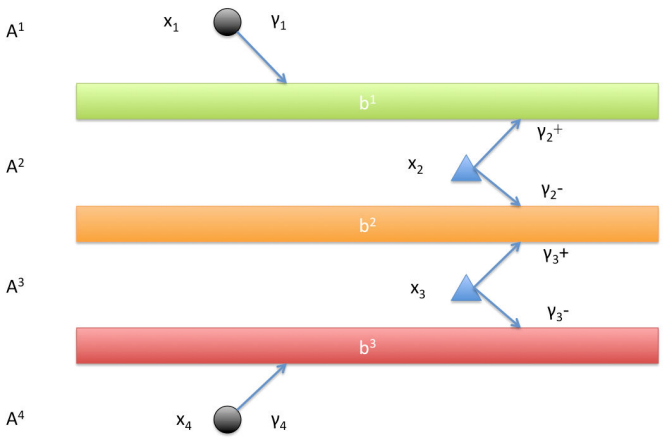

Figure 3-2: Representation of constraints relaxation algorithm

Inconsistency resolution via constraints relaxation

Based on the algorithm presented in the previous subsection, a subset of maximal cardinality that can be represented by an MR-sort model is obtained. At the same time, its complementary set is *deleted*. However, in order to help the DMs understand in what way the identified inconsistent inputs conflict with the others; and guide them to reconsider and possibly modify their judgments, a constraints relaxation algorithm is here proposed.

As presented in previous situation, each alternative x_p can provide one or two constraints with respect to its assignment. As presented in Figure 3-2, we introduce the following binary variables: γ_p, for the alternatives originally assigned to extreme categories, i.e., A^1 and A^4; γ_p^+ and γ_p^- for the alternatives originally assigned to intermediate categories, i.e., A^2 and A^3: in particular, γ_p^+ refers to the fulfillment of the constraint associated to the better category lower profiles, whereas γ_p^- refers to the fulfillment of the constraint associated to the worse category lower profiles.

As in the previous case, the algorithm identifies a subset $D_{TR}^* \subseteq D_{TR}$ of maximal cardinality that can generate an MR-sort model with proper formulation. In addition, for each of the alternatives that are not accepted into the subset D_{TR}^*, the corresponding inconsistent constraints are also targeted: for example, if for one alternative x_p we obtain $\gamma_p^+ = 0$ (resp., $\gamma_p^- = 0$), then this alternative should be classified in a bet-

ter (resp., worse) category; in other words, its original assignment is underestimated (resp., overestimated). The same criterion is applied to the alternatives that are originally assigned to the best or worst category. More details of the inconsistency study are presented in appended Paper (iii).

3.3 Performance assessment of the MR-Sort classification model for the vulnerability assessment

Due to the finite (typically small) size of the set of training classification examples usually available in the analysis of real complex safety-critical systems, the performance of the classification model is impaired. In particular, (i) the classification *accuracy* (resp., error), i.e., the expected fraction of patterns correctly (resp., incorrectly) classified, is typically reduced (resp., increased); (ii) the classification process is characterised by significant uncertainty, which affects the *confidence* of the classification-based vulnerability model: in this thesis work, the confidence in a classification assignment is defined as in [10], i.e., as the probability that the class assigned by the model to a given (single) pattern is the correct one. Obviously, there is the possibility that a classification model assigns correctly a very large (expected) fraction of patterns (i.e., the model is very accurate), but at the same time *each* (correct) assignment is affected by significant uncertainty (i.e., it is characterised by low confidence). It is worth mentioning that besides the scarcity of training data, there are many additional sources of uncertainty in classification problems (e.g., the accuracy of the data, the suitability of the classification technique used, etc.): however, they are not considered in this work.

The performance of the classification model (i.e., the classification accuracy - resp., error - and the confidence in the classification) needs to be quantified: this is of paramount importance for taking robust decisions in the vulnerability analyses of safety-critical systems [8][89][48].

In the present work, three different approaches are used to assess the performance

of a classification-based MR-Sort vulnerability model in the presence of small training data sets. The first is a model-retrieval based approach [82], which is used to assess the expected percentage error in assigning new alternatives. The second is based on *bootstrapping* the available training set in order to build an ensemble of vulnerability models [102][35]; the method can be used to assess both the accuracy and the confidence of the model: in particular, the confidence in the assignment of a given alternative is given in terms of the full (probability) distribution of the possible vulnerability classes for that alternative (built on the bootstrapped ensemble of vulnerability models) [10]. The third is based on the Leave-One-Out Cross-Validation (LOOCV) technique, in which one element of the available data set is (left out and) used to test the accuracy of the classification model built on the remaining data: also this approach is employed to estimate the accuracy of the classification vulnerability model as the expected percentage error, i.e., the fraction of alternatives incorrectly assigned (computed as an average over the left-out data).

3.3.1 Model-retrieval based approach

The first method is based on the model-retrieval approach proposed in [82]. A fictitious set D_{TR}^{rand} of N alternatives $\{x_p^{rand} : p = 1, 2, ..., N\}$ is generated by random sampling within the ranges X_i of the criteria, $i = 1, 2, ..., n$. Notice that the size N of the fictitious set D_{TR}^{rand} has to be the same as the real training set D_{TR} available, for the comparison to be fair. Also, a MR-Sort classification model $M(\cdot|\omega^{rand}, b^{rand})$ is constructed by randomly sampling possible values of the internal parameters, $\{\omega_i : i = 1, 2, ..., n\}$ and $\{b^h : h = 1, 2, ..., M-1\}$. Then, we simulate the behaviour of a Decision Maker (DM) by letting the (random) model $M(\cdot|\omega^{rand}, b^{rand})$ assign the (randomly generated) alternatives $\{x_p^{rand} : p = 1, 2, ..., N\}$. In other words, we construct a learning set D_{TR}^{rand} by assigning the (randomly generated) alternatives using the (randomly generated) MR-Sort model, i.e., $D_{TR}^{rand} = \{(x_p^{rand}, \Gamma_p^M) : p = 1, 2, ..., N\}$, where Γ_p^M is the class assigned by model $M(\cdot|\omega^{rand}, b^{rand})$ to alternative x_p^{rand}, i.e., $\Gamma_p^M = M(x_p^{rand}|\omega^{rand}, b^{rand})$. Subsequently, a new MR-Sort model $M'(\cdot|\omega', b')$, compatible with the training set D_{TR}^{rand}, is inferred using the MIP for-

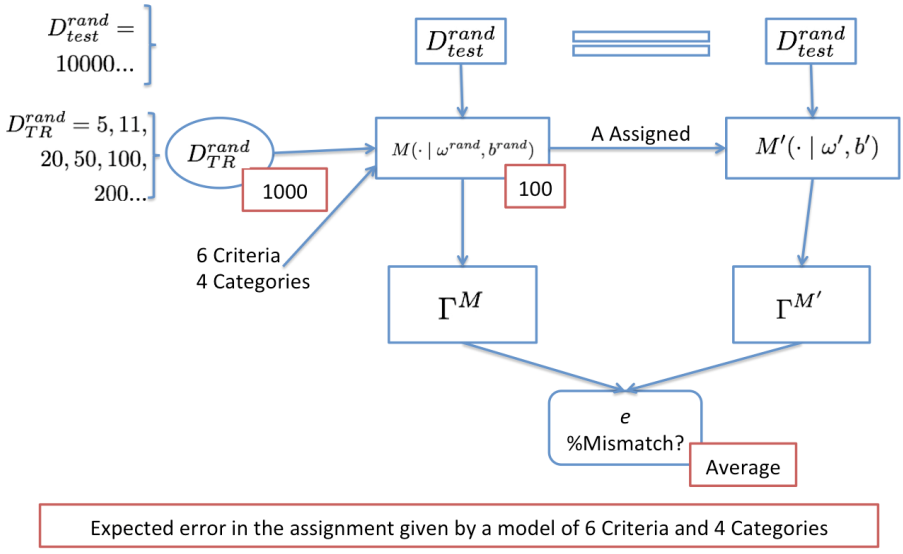

Figure 3-3: The general structure of the model-retrieval approach

mulation summarised in Chapter 3.2.1 and in the appended Paper (ii). Although models $M(\cdot|\omega^{rand}, b^{rand})$ and $M'(\cdot|\omega', b')$ may be quite different, they coincide on the way they assign elements of D^{rand}_{TR}, by construction. In order to compare models M and M', we randomly generate a (typically large) set D^{rand}_{test} of *new* alternatives $D^{rand}_{test} = \{x^{test,rand}_p : p = 1, 2, ..., N_{Test}\}$ and we compute the percentage of assignment errors, i.e., the proportion of these N_{Test} alternatives that models M and M' assign to different categories.

In order to account for the randomness in the generation of the training set D^{rand}_{TR} and of the model $M(\cdot|\omega^{rand}, b^{rand})$, and to provide robust estimates for the assignment errors ϵ, the procedure outlined above is repeated for a large number N_{sets} of random training sets $D^{rand,j}_{TR}, j = 1, 2, ..., N_{sets}$; in addition, for each set j the procedure is repeated for different random models $M(\cdot|\omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models}$. The sequence of assignment errors thereby generated, $e_{jl}, j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$, is then averaged to obtain a robust estimate for ϵ. The procedure is sketched in Figure 3-3.

Notice that this method does not make any use of the original training set D_{TR} (i.e., of the training set constituted by real-world classification examples). In this view, the model retrieval-based approach can be interpreted as a tool to obtain an abso-

lute evaluation of the expected error that an "average" MR-Sort classification model $M(\cdot|\omega, b)$ with M categories, n criteria and trained by means of an "average" data set of given size N makes in the task of classifying a new generic (unknown) alternative.

3.3.2 Bootstrap method

A way to assess *both* the accuracy (i.e., the expected fraction of alternatives correctly classified) *and* the confidence of the classification model (i.e., the probability that the category assigned to a given alternative is the correct one) is by resorting to the bootstrap method [35], which is used to create an ensemble of classification models constructed on different data sets bootstrapped from the original one [137]: the final class assignment provided by the ensemble is based on the combination of the individual output of classes provided by the ensemble of models [10].

The basic idea is to generate different training datasets by random sampling with replacement from the original one [35]: such different training sets are used to build different individual classification models of the ensemble. In this way, the individual classifiers of the ensemble possibly perform well in different regions of the training space and thus they are expected to make errors on alternatives with different characteristics; these errors are balanced out in the combination, so that the performance of the ensemble of bootstrapped classification models is in general superior than that of the single classifiers [137][24]. This is a desirable property since it is a more realistic simulation of the real-life experiment from which our dataset was obtained. In this work, the output classes of the single classifiers are combined by *majority voting*: the class chosen by most classifiers is the ensemble assignment. Finally, the accuracy of the model is given by the fraction of the patterns correctly classified. The bootstrap-based empirical distribution of the assignments given by the different classification models of the ensemble is then used to measure the confidence in the classification of a given alternative x that represents the probability that this alternative is correctly assigned [10][11]. In more detail, the main steps of the bootstrap algorithm are as follows (Figure 3-4):

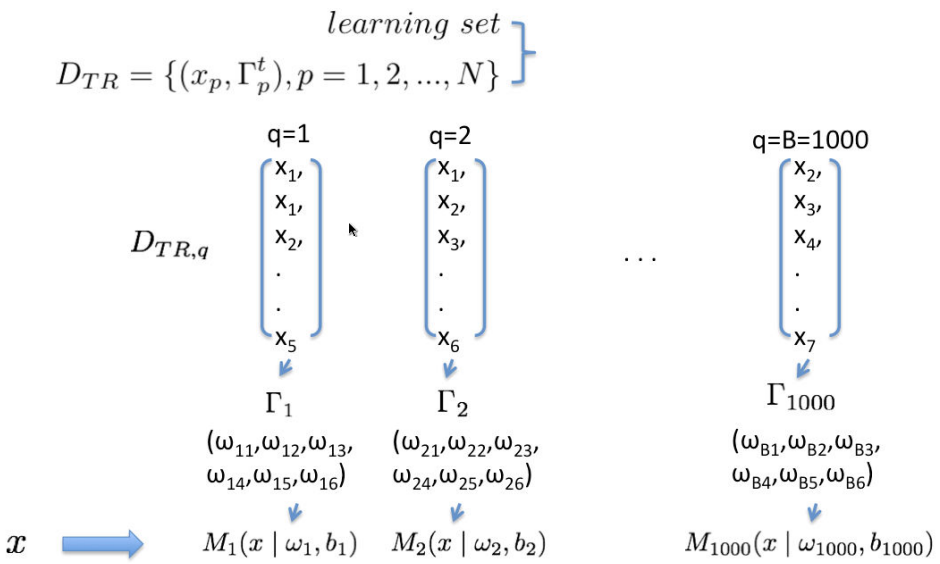

Figure 3-4: The bootstrap algorithm

1. Build an ensemble of B (typically of the order of $500-1000$) classification models $\{M_q(\cdot|(\omega_q, b_q) : q = 1, 2, ..., B)\}$ by random sampling with replacement from the original data set D_{TR} and use each of the bootstrapped models $M_q(\cdot|\omega_q, b_q)$ to assign a class $\Gamma_p^q, q = 1, 2, ..., B$, to a given alternative x_p of interest (notice that Γ_p^q takes a value in $A^h, h = 1, 2, ..., M$). By so doing, a bootstrap-based empirical probability distribution $P(A^h|x_p), h = 1, 2, ..., M$ for category A^h of alternative x_p is produced, which is the basis for assessing the confidence in the assignment of alternative x_p. In particular, repeat the following steps for $q = 1, 2, ..., B$:

   a. Generate a bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$, by performing random sampling with replacement from the original data set $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$ of N input/output patterns. The data set $D_{TR,q}$ is thus constituted by the same number N of input/output patterns drawn among those in D_{TR}, although due to the sampling with replacement some of the patterns in D_{TR} will appear more than once in $D_{TR,q}$, whereas some will not appear at all.

   b. Build a classification model $\{M_q(\cdot|\omega_q, b_q) : q = 1, 2, ..., B\}$, on the basis of

the bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$.

c. Use the classification model $M_q(\cdot|\omega_q, b_q)$ to provide a class $\Gamma_p^q, q = 1, 2, ..., B$ to a given alternative of interest, i.e., $\Gamma_p^q = M_q(x_p|\omega_q, b_q)$.

2. Combine the output classes $\Gamma^q, q = 1, 2, ..., B$ of the individual classifiers by majority voting: the class chosen by most classifiers is the ensemble assignment $\Gamma_p^{ens}, i.e.,\ \Gamma_p^{ens} = argmax_{A^h}[card_q\{\Gamma_p^q = A^h\}]$.

3. As an estimation of the confidence in the majority-voting assignment Γ_p^{ens}(step 2, above), we consider the bootstrap-based empirical probability distribution $P(A^h|x_p), h = \{1, 2, ..., M\}$, i.e., the probability that category A^h is the correct category given that the (test) alternative is x_p [82]. The estimator of $P(A^h|x_p)$ here employed is: $P(A^h|x_p) = \frac{\sum_{q=1}^{B} I\{\Gamma^q = A^h\}}{B}$, where $I\{\Gamma^q = A^h\} = 1$, if $\Gamma^q = A^h$, and 0 otherwise.

4. Finally, the error of classification is presented by the fraction of the number of the alternatives being assigned by the classification model and the total number of the alternatives. The accuracy of the classification model is defined as the complement to 1 to the error.

3.3.3 Cross validation

Leave-One-Out Cross-Validation (LOOCV) is a particular case of the cross-validation method. In cross-validation, the original training set D_{TR} is divided into N partitions, $Pt_1, Pt_2, ..., Pt_N$, and the elements in each of the partitions are classified by a model trained by means of the elements in the remaining partitions (Leave-p-out Cross-Validation) [11]. The cross-validation error is, then, the average of the N individual error estimates. When N is equal to the number of elements N in D_{TR}, the result is leave-one-out cross-validation (LOOCV), in which each instance $x_p, p = 1, 2, ..., N$ is classified by all the instances in D_{TR} except for itself [129]. For each instance $x_p, p = 1, 2, ..., N$ in D_{TR}, the classification accuracy is 1 if the element is classified correctly and 0 if it is not. Thus, the average LOOCV error (resp. accuracy) over

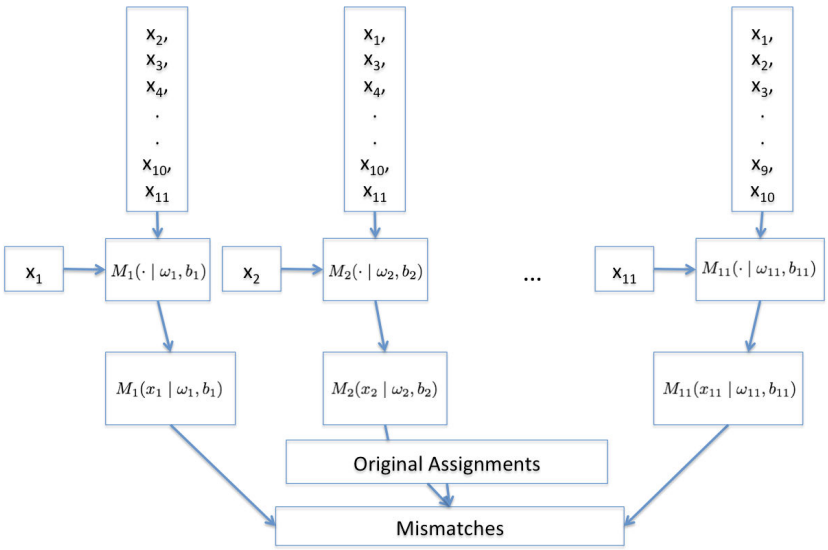

Figure 3-5: Leave-one-out cross-validation study procedure

all the N instances in D_{TR} is ϵ/N (resp. $1-\epsilon/N$), where ϵ (resp. $N-\epsilon$) is the number of elements incorrectly (resp. correctly) classified. Thus, the accuracy in the assignment is estimated as $1-\epsilon/N$. With respect to the leave-p-out cross-validation, the leave-one-out cross-validation (LOOCV) produces a smaller bias of the true error rate estimator. However, the computational time increases significantly with the size of the data set available. This is the reason why the LOOCV is particularly useful in the case of small data sets. In addition, for *very sparse* datasets (e.g., of size lower than or equal to ten), we may be "*forced*" to use LOOCV in order to maximise the number of training examples employed and to generate training sets containing an amount of information that is sufficient and reasonable for building an empirical model [44]. In Figure 3-5, the algorithm is sketched with reference to a training set D_{TR} containing $N = 11$ data (as in the application considered in Chapter 5).

Chapter 4

INVERSE MULTICRITERIA CLASSIFICATION PROBLEM: IDENTIFYING PROTECTIVE ACTIONS TO REDUCE VULNERABILITY

What constitutes a "multiple criteria decision making (MCDM)" problem? Clearly there must be some decision to be made! Such decision may constitute a simple choice between two or more (perhaps even infinitely many) well-defined alternatives. The problem is then simply that of making the "best" choice in some sense (as presented in Chapter 2 and 3). At the other extreme, there may be a vague sense of unease that we (personally or corporately) need to "do something" about a situation which is found unsatisfactory in some way. The decision problem then constitutes much more than simply the evaluation and comparison of alternatives. It involves also an in-depth consideration of what it is that is "unsatisfactory", and the creative generation of possible courses of actions to address the situation [13].

The classification problem has been widely studied in the literature because of its ap-

plicability to a wide variety of problems [34][60][4][22][16][23][40][42][62], as presented previously (Chapter 3). Based on the classification model and results obtained in the previous Chapter, the problem is structured as an inverse multicriteira classification problem [3][1][83]. The idea is that changes are made in the independent variables of a sample so that the sample can be classified into a more desirable class [97][86][2]. In other words, we determine the features to be used to create a record which will result in a *desired* class label. Particularly, in this work we aim to identify a set of protective actions to reduce the vulnerability of a (group of) safety-critical system(s) eventually under budget limitations. This system can be used for a variety of decision support applications which have predetermined task criteria.

4.1 Inverse classification problem: general framework

In the inverse classification problem, we would like to determine the action oriented feature variables for an *incompletely specified* test data set. Typically, these feature variables are decision variables for an optimization or decision support application. The aim is to decide the choice of the actions so that these feature variables are modified in such a way so as that the resulting records would belong to a set of desired class variable values for the test data set.

If there is no limitation on the choices of the actions (e.g., number of actions that can be applied) the problem can be formulated as that for the case of the training data set, both the feature and class variables are completely defined in it. On the other hand, for the case of the test data set, the class variables are completely defined but the feature variables are not. Thus, each test data example has a *desired class label* associated with it. The aim of the inverse classification problem is to choose the test feature variables such that the corresponding classification accuracy with respect to the *desired* test classes is maximized.

If there are action related constraints (e.g., number of actions can be applied at the same time, budget limitation etc.) then the problem is modified. The class variables and the feature variables are all not definitively defined. Under the given constraints

concerning the choice of the actions, the aim of the inverse classification problem is to choose the "optimal" set of actions that can modify the feature variables such that the corresponding class variables are brought to a possibly "desired" class label.
The problem is "inverse" because the usual mapping is from a case to its unknown category. The increased application of classification systems in business suggests that this inverse problem can be of significant benefit to decision-makers as a form of sensitivity analysis.
We note that the inverse classification problem is different from the classification or imputation problem on missing data sets. In the classification problem on missing data, we try to determine the *unknown* class variable with incompletely defined features. On the other hand, in the inverse classification problem, we try to determine the *action-oriented* missing variables in order to achieve a *desired result* which is reflected in the class variable. The inverse classification problem is useful for a number of action-driven applications in which the features can be used to define certain actions which drive the decision support system towards a *desired* end-result [2].

4.2 Inverse classification problem: framework proposed in the present thesis

To illustrate the methodology, we carry on the definition of the classification model of the safety-critical systems presented in the previous Chapter: we consider a set of N NPPs ($x_p, p \in \{1, 2, ..., N\}$) characterized by m = 16 basic features ($crit^j, j \in \{1, 2, ..., m\}$). On the basis of these $m = 16$ features, the NPPs are assigned to $M = 4$ pre-defined categories ($A^h \in \{1, 2, ..., M\}, h \in \{1, 2, ..., N\}$), where A^1 represents the best situation, i.e., lowest vulnerability. Let $act = \{act^1, act^2, ..., act^k, ..., act^F\}$ denote the available set of actions, each of which can influence on one or more basic criteria $crit^j, j \in \{1, 2, ..., m\}$ (Figure 4-1) with different intensity, as measured by a set of coefficients $\{coeff^{kj}, k \in \{1, 2, ..., F\}, j \in \{1, 2, ..., m\}\}$. In other words, $coeff^{kj}$ is

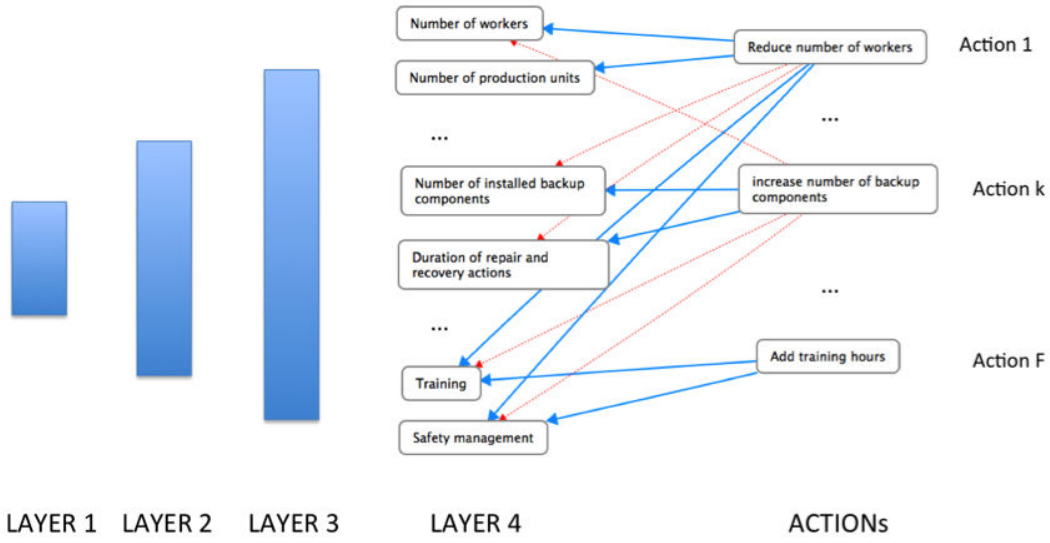

Figure 4-1: Schema of direct actions for basic criteria

the "weight" of the influence of action k on attribute j (the higher the absolute value of $coeff^{kj}$, the stronger the effect of action k on attribute j). Notice that a positive (resp. negative) coefficient $coeff^{kj}$ means that action k has an ameliorative (resp. deteriorative) effect on attribute j, whereas if $coeff^{kj}$ is equal to zero, then criterion j is not influenced by action k. The implementation of one or more actions modifies the attribute values $crit^j, j \in \{1, 2, ..., m\}$ and as a result, the vulnerability of the system (i.e., the assignment by the classification model) may change. In this work, we assume that the total effect of the available set of actions $act = \{act^1, act^2, ..., act^k, ..., act^F\}$ on criterion j is obtained by a linear superposition of the effects of each action act^k:

$$crit'^j = crit^j + \sum_{k=1}^{F} coeff^{kj} * act^k, k \in \{1, 2, ..., F\}, j \in \{1, 2, ..., m\}. \quad (4.1)$$

where $crit'^j$ is the value of attribute j after the identified set of available actions has been implemented.

Also, let $Cost(x_p, act'), act' \subseteq act$ denote the cost of the combination of actions act' applied to x_p. If $c_k^p (p \in \{1, 2, ..., N\}, k \in \{1, 2, ..., F\})$ is the cost of action k on x_p,

then:

$$Cost(x_p, act') = \sum_k c_k^p, k \in \{1, 2, ..., F\}. \tag{4.2}$$

The inverse classification problem can, then, be formulated as follows: given a limited budget B_g for the entire group of NPPs considered, find out for each of the NPPs the best combination of actions that provide the maximal possible reduction in their vulnerability level $A_p^{\lambda'}, A_p^{\lambda'} \in \{1, 2, 3, 4\}, p \in \{1, 2, ..., N\}$ (as presented in Chapter 2, the smaller the category value, the less vulnerable the NPP) under budget constraint. In particular, we have chosen the strategy to reduce, under budget constraint, the global vulnerability of a group of alternatives in giving priority to the NPPs that are originally assigned to the worst category, in other words, we try to maximize a properly weighted sum of the ameliorations in the vulnerability categories undergone by all the NPPs. This is mathematically represented by the objective function I^x as an intermediate value of the objective function in considering a set of alternatives $x = \{x_p, p \in \{1, 2, ..., N\}\}$:

$$I^x = \rho_3 * Q_{43} + \rho_2 * Q_{32} + \rho_1 * Q_{21} \tag{4.3}$$

where $Q_{n(n-1)(n\in\mathbb{Z})}$ represents the number of NPPs among the N available ones $\{x|x_p, p \in \{1, 2, ..., N\}\}$ that are ameliorated from category A^n to category $A^{(n-1)}$ by a given combination of actions. The constants $\{\rho_i|i \in \{1, 2, 3\}\}$ represent weights that we assigned to the number of ameliorated NPPs $Q_{n(n-1)(n\in\mathbb{Z})}$. We assign the following set of weights:

$$\rho_3 = 100, \rho_2 = 50, \rho_1 = 25. \tag{4.4}$$

In this case, by maximizing the objective function I^x, high importance is given to the amelioration of the worst (most vulnerable) NPPs.

To address the inverse classification problem, we first adopt a pragmatic approach based on sensitivity analysis [110][55][111], introducing indicators that quantify the

variation in the vulnerability class that a safety-critical system is expected to undergo upon implementation of a given set of actions. Then, three different optimization approaches will be explored: (i) one single classification model is built to evaluate and minimize system vulnerability; (ii) an ensemble of compatible classification models, generated by the bootstrap method, is employed to perform a "robust" optimization, by considering the "worst-case" scenario; (iii) finally, a distribution of classification models, still obtained by bootstrap, is considered to address vulnerability reduction in a "probabilistic" fashion.

4.3 Solution to the inverse classification problem: sensitivity indicators

All multicriteria methods call for the identification of the key factors which will form the basis of an evaluation. These are referred to variously as: values, (fundamental) objectives, criteria, (fundamental) point of view. The extent to which, and the way in which, these key factors are elaborated in the model structure differs between the methodologies [13]. Research in classification analysis has focused on developing models/algorithms for correct classification of training and holdout sample data (as presented in Chapter 3). Most classification systems lack the ability to systematically conduct sensitivity analysis [86]. In this thesis work, we have developed a process to analyze the different actions based on several novel defined indicators as presented in the following paragraphs:

We consider the group of N' vulnerability-class labeled known (available) safety-critical systems (NPPs) used to train the MR-Sort classification model and study the sensitivity of their categories of vulnerability to the implementation of the available protective actions. We denote the original categories of these NPPs as $A^h, A^h \in \{1, 2, ..., M\}, h \in \{1, 2, ..., N'\}$ and the new categories resulting from the application of a set of protective actions as $A^h_\lambda, A^h_\lambda \in \{1, 2, ..., M\}, h \in \{1, 2, ..., N'\}$.

Let $N\uparrow$ be the number of NPPs that are improved after the action(s):

$$N\uparrow = \sum_{h=1}^{N'} C_h, h \in \mathbb{N}, \tag{4.5}$$

$$C_h = 1,\ if\ A^h > A^h_\lambda, \tag{4.6}$$

$$C_h = 0,\ if\ A^h < A^h_\lambda. \tag{4.7}$$

Then, $\frac{N\uparrow}{N'}$ can be interpreted as an estimate of the percentage of new (i.e., different from the ones of the training set) NPPs that can be expected to be improved after such action(s) is (are) implemented on them.

Dually, $N\downarrow$, is the number of NPPs that are expected to be deteriorated after the action(s):

$$N\downarrow = \sum_{h=1}^{N'} C_h, h \in \mathbb{N}, \tag{4.8}$$

$$C_h = 1,\ if\ A^h < A^h_\lambda, \tag{4.9}$$

$$C_h = 0,\ if\ A^h > A^h_\lambda. \tag{4.10}$$

Notice that a "deterioration" (i.e., an increase in the vulnerability category) is possible because some of the actions may have positive effects on some subcriteria but negative effects on some others (cf Section 3). Then, $\frac{N\downarrow}{N'}$ can be interpreted as an estimate of the percentage of new NPPs (i.e., different from the ones of the training set) that can be expected to be deteriorated after such action(s) is (are) implemented on them.

We consider the quantity $\Delta N = \frac{N\uparrow}{N'} - \frac{N\downarrow}{N'}$ to combine the effects of both positive and negative influences of the actions in the expected "net" amount of ameliorated NPPs. Considering that the evaluation framework is based on $M = 4$ categories, it seems reasonable to consider not only the number of NPPs that are ameliorated or deteriorated, but also the amount of variation in category of vulnerability of each of them. To this aim, we introduce the following indicators to combine the amount of

variation in vulnerability with the number of NPPs whose vulnerability category has changed after the actions.

In particular, $\Delta M \uparrow$ is defined as the total variation of category underwent by the *ameliorated* NPPs:

$$\Delta M \uparrow = \sum_{h=1}^{N'} (M_h * C_h), h \in \mathbb{N}, \tag{4.11}$$

$$M_h = A^h - A^h_\lambda, \tag{4.12}$$

$$C_h = 1, if\ A^h > A^h_\lambda, \tag{4.13}$$

$$C_h = 0, if\ A^h < A^h_\lambda. \tag{4.14}$$

Thus, $\frac{\Delta M \uparrow}{N'}$ can be interpreted as the variation in vulnerability category that a new ameliorated plant is expected to undergo when the chosen combination of actions is applied.

Dually, $\Delta M \downarrow$ is defined as:

$$\Delta M \downarrow = \sum_{h=1}^{N'} (M_h * C_h), h \in \mathbb{N}, \tag{4.15}$$

$$M_h = A^h - A^h_\lambda, \tag{4.16}$$

$$C_h = 1, if\ A^h < A^h_\lambda, \tag{4.17}$$

$$C_h = 0, if\ A^h > A^h_\lambda. \tag{4.18}$$

Thus, $\frac{\Delta M \downarrow}{N'}$ can be seen as the variation in vulnerability category that a new *deteriorated* plant is expected to undergo when the chosen combination of actions is applied. Finally, $\overline{\Delta M} = \frac{\Delta M \uparrow}{N'} - \frac{\Delta M \downarrow}{N'}$ combines the effects of both positive and negative influences of the actions and it can be seen as the "net" variation in vulnerability category that a newly analyzed NPP is expected to undergo after the application of the given set of actions. The net expected variation in vulnerability category $\overline{\Delta M}$ quantifies the influence of the actions upon the NPPs. However, this measure does not take into

account the original category assignment of the NPPs: for example, in practice there is a difference between taking a NPP from category 4 to 3 and taking it from 2 to 1, even if the category variation is 1 in both cases. To consider this, we introduce the indicator $\Delta S \uparrow$, defined as the ratio between the sums of the variations of vulnerability category underwent by the ameliorated NPPs and the sum of the corresponding maximum possible category variations (i.e., the sum of the category variations that the NPPs would undergo if they were ameliorated to the best possible vulnerability category):

$$\Delta S \uparrow = \frac{\Delta M \uparrow}{E}, \tag{4.19}$$

$$E = \sum_{h=1}^{N'} (A^h - A^{best}) * C_h, h \in \mathbb{N}, \tag{4.20}$$

$$C_h = 1, if\ A^h > A^h_\lambda, \tag{4.21}$$

$$C_h = 0, if\ A^h < A^h_\lambda. \tag{4.22}$$

The indicator $\Delta S \uparrow$ quantifies the influence of the actions on NPPs, relative to their original categories: the lower $\Delta S \uparrow$ is, the higher the influence of the chosen set of actions is on the NPPs originally of a relatively low category.
Based on the above indicators, an algorithm is proposed to rank different combinations of actions according to their effectiveness in reducing the vulnerability of safety-critical systems. The algorithm proceeds as follows:

1. Rank the (combinations of) actions according to the value of $\overline{\Delta M}$ (the higher the value of $\overline{\Delta M}$, the more effective the combination of actions in reducing vulnerability):

    - combinations of actions that have a negative value of $\overline{\Delta M}$ ($\overline{\Delta M} < 0$) are expected to increase the vulnerability of a NPP: this is due to the fact that some actions may have a deteriorated effect on some of the subcriteria that more than counter balances the positive effects on their subcriteria. The identification of the combination of actions with $\overline{\Delta M} < 0$ allows the

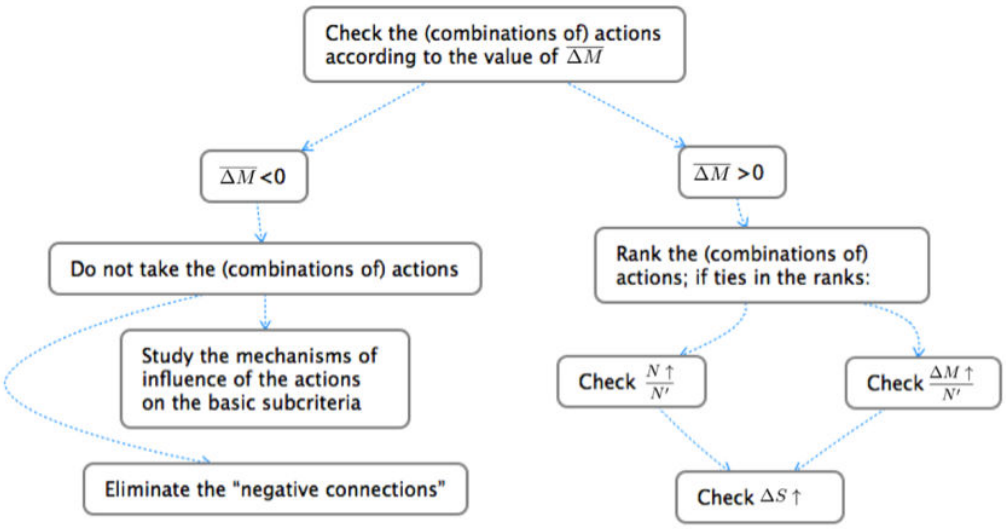

Figure 4-2: Schema of decision logic for selecting an action

analyst to (i) study the mechanisms of influence of the actions on the basic subcriteria (Layer 4 in Figures 2-1 and 4-1) and (ii) if possible, eliminate the "negative connections", i.e., the negative dependencies between some actions and some criteria (e.g., by identifying alternative actions for dealing with these "critical" subcriteria).

- the actions that have a positive value of $\overline{\Delta M}$ ($\overline{\Delta M} > 0$) are expected to reduce the vulnerability and are assigned higher rankings (the higher $\overline{\Delta M}$, the higher the ranking).

2. If several combinations of actions have the same value of $\overline{\Delta M}$, then consider the other indicators (i.e., $\frac{N\uparrow}{N'}$ and $\frac{\Delta M\uparrow}{N'}$): depending on the judgment of the DMs, higher importance may be given to those actions that produce a larger expected number of improved NPPs ($\frac{N\uparrow}{N'}$) or to those that generate a higher "expected class improvement" ($\frac{\Delta M\uparrow}{N'}$).

3. If some combinations still have the same ranking, analyze indicator $\Delta S\uparrow$ to check which actions have stronger impact on the NPPs of low categories.

4.4 Solution to the inverse classification problem: optimization

As presented in the previous section, the choice of the protective actions is studied based on the sensitivity indicators. In this section, we adopt an optimization-based decision-making approach solved by CPLEX solver. The study is carried in three different circumstances, named simple optimization, robust optimization and probabilistic optimization as presented in the following subsections. These optimizations show three different views of how the choices of the protective actions for each of the NPPs could be made in taking the classification evaluation model into account. The mathematical formulation of the inverse problem is also adapted.

4.4.1 Simple optimization

As presented in Chapter 3, and in more details in [125], we are able to obtain a compatible classification model as $M^*(\cdot|\omega^*, b^*)$ (with ω^* the weights and b^* the lower profiles) based on all the pre-assigned alternatives in the given set D_{TR} through a disaggregation process. We name this model the "*optimum*" compatible classification model. The optimization process aims at finding an optimal set of actions for each of the NPPs for which the objective function I^x is maximized: this will improve the performance of the group of NPPs, giving priority to the worst ones. In more detail, the problem can be formulated as follows:

$$Find\ act'_p = arg\ Max_{\{act'_p, p=1,2,...,N\}}(I^x(act'_p, M^*)), \tag{4.23}$$

$$s.t.\ \sum_p Cost(x_p, act'_p) \leq B_g, \tag{4.24}$$

$$\{x|x_p, p \in \{1, 2, ..., N\}\}. \tag{4.25}$$

Under the constraint of budget limitation, we find the combination of protective actions that maximize the value of the objective function I^x, presented above.

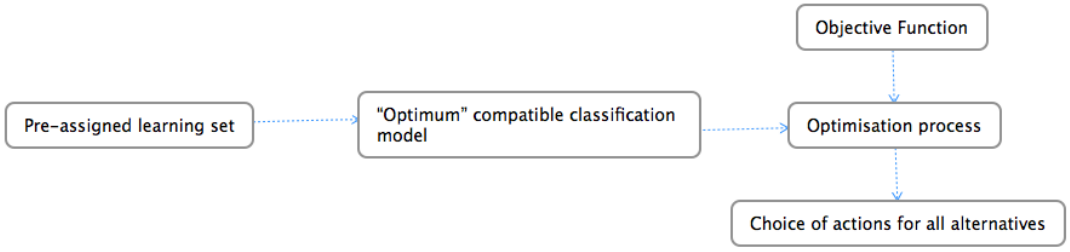

Figure 4-3: Representation of Simple optimization

4.4.2 Robust optimization

The optimization approach introduced above provides a choice of protective actions for the NPPs using (only) the "*optimum*" classification model. However, for the training set of pre-assigned alternatives there are a number of compatible classification models. We aim at finding out the combination of protective actions (for each of the NPPs) that can ameliorate the NPPs to a satisfactorily low level of vulnerability, considering all compatible classification models. In other words, the combination of actions that we obtain should be "*robust*" to the (model) vulnerability assessing from the fact that the empirical classification model is trained with a finite set of data and, thus, multiple models are compatible.

To this aim, the bootstrap method [35] is applied to create an ensemble of classification models constructed on different data sets bootstrapped from the original one [137].

The basic idea is to generate different training datasets by random sampling with replacement from the original one [35]: such different training sets are used to build different individual classification models of the ensemble. In this way, the individual classifiers of the ensemble possibly perform well in different regions of the training space.

In more detail, the main steps of the bootstrap algorithm are as follows (Figure 4-4): Build an ensemble of B classification models $\{M_q(\cdot|(\omega_q, b_q) : q = 1, 2, ..., B)\}$ by random sampling with replacement from the original data set $D_{TR}(x_p, p \in \{1, 2, ..., N\})$ and integrate each of the bootstrapped models $M_q(\cdot|\omega_q, b_q)$ into the optimization pro-

gram which later allow us to assign a class $\Gamma_p^q, q = 1, 2, ..., B$, to a given alternative x_p of interest (notice that Γ_p^q takes a value in $\{A^h : h = 1, 2, 3, 4\}$).

a. Generate a bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$, by performing random sampling with replacement from the original data set $D_{TR} = \{x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$ of N input/output patterns. The data set $D_{TR,q}$ is thus constituted by the same number N of input/output patterns drawn among those in D_{TR}, although due to the sampling with replacement some of the patterns in D_{TR} will appear more than once in $D_{TR,q}$, whereas some others will not appear at all.

b. Build a classification model $\{M_q(\cdot|\omega_q, b_q) : q = 1, 2, ..., B\}$, on the basis of the bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$.

Given the bootstrapped ensemble, the mathematical formulation of the robust optimization is as follows:

$$Find\ act'_p = arg\ Max_{\{act'_p, p=1,2,...,N\}}\ Min_q(I^x(act'_p, M_q)), \tag{4.26}$$

$$s.t.\ \sum_p Cost(x_p, act'_p) \leq B_g, \tag{4.27}$$

$$\{x|x_p, p \in \{1, 2, ..., N\}\}, \tag{4.28}$$

$$\{M|M_q \in M, q \in \{1, 2, ..., B\}\}. \tag{4.29}$$

A large number B (=100) of compatible classification models $\{M|M_q \in M, q \in \{1, 2, ..., B\}\}$ are typically generated by bootstrap. Correspondingly, the minimum value $Min_M(I^x(act'_i, M_q))$ of objective function $I^x(act'_p, M_q)$ over the B compatible models in correspondence of each set of actions can be gathered. In particular, a distribution of vulnerability classes can be obtained for each NPP. Then, based on the distribution and applying the majority-voting rule, we assign each NPP to its most likely after-action category. Then, the optimization solver aims at finding the optimal combination of actions that robustly and conservatively maximize the worst value of the objective function $I^x(act'_p, M_q)$.

In more detail, the robust optimization algorithm proceeds as follows:

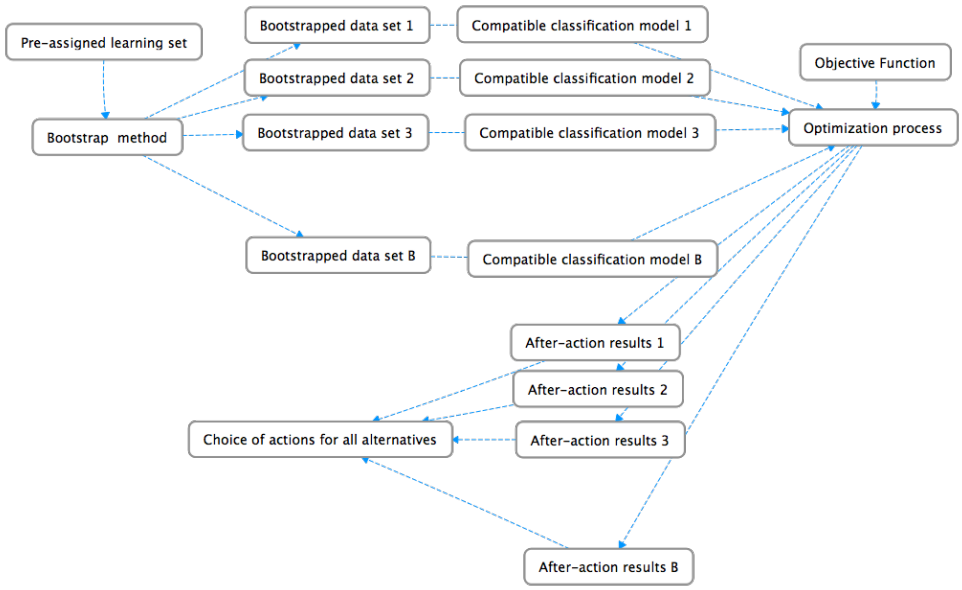

Figure 4-4: Representation of Robust optimization

1. The solver proposes a set of actions for each x_p; each bootstrapped classification moedel $M_q(\cdot|\omega_q, b_q)$ is used to provide an after-action class $\Gamma_p^q, q = 1, 2, ..., B$ to a given alternative of interest, i.e., $\Gamma_p^q = M_q(x_p|\omega_q, b_q)$;

2. On the basis of the results obtained at step 1 above, a value of function $I^x(act'_p, M_q)$ is computed for each compatible model $M_q(\cdot|\omega_q, b_q), q = 1, 2, ..., B$, to obtain an ensemble of values of $I^x(act'_p, M_q)$;

3. The minimum (i.e., worst) value among $I^x(act'_p, M_q), q = 1, 2, ..., B$, is taken as the objective function to maximize; in other words, we aim at identifying the set of actions able to improve the "worst-case scenario" over the possible compatible models;

4. We repeat the steps above for different combinations of actions $act'_p, p = 1, 2, ..., N$ in order to find out the combination of actions for each of the considered NPPs that can ameliorate the worst after action result to the possibly best situation.

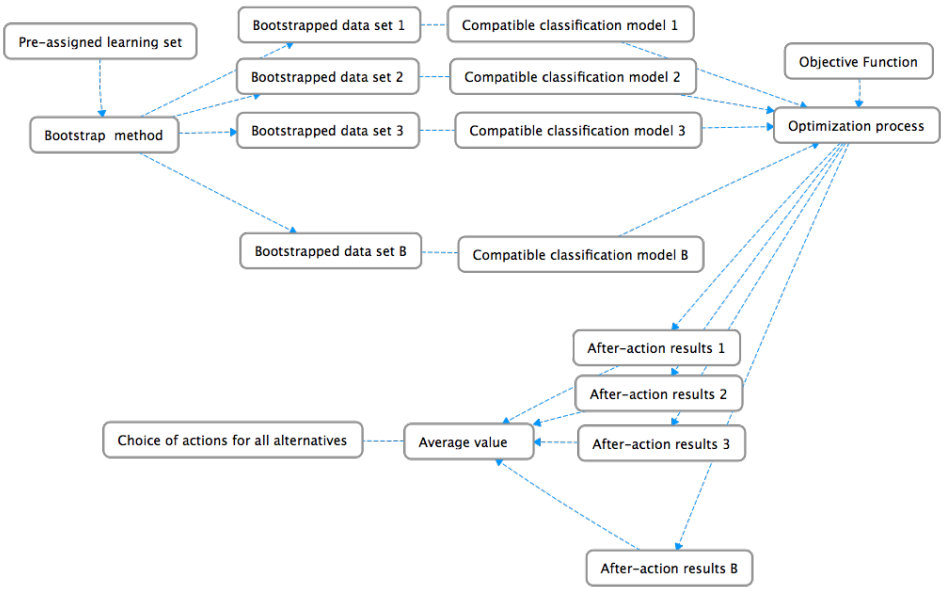

Figure 4-5: Representation of Probabilistic optimization

4.4.3 Probabilistic optimization

The main steps (Figure 4-5) are the same as those of the Robust optimization as presented in Figure 4-4, but the objective function is changed. Instead of improving the worst case over all the models, we choose to improve the expected value of the probablity distribution of the function I^x. Thus, in this case, we "ignore" some of the "extreme" classification models generated by bootstrap.

The mathematical formulation of the problem is as follows:

$$Find\ act'_p = arg\ Max_{\{act'_p, p=1,2,...,N\}} \frac{1}{B}\sum_{q=1}^{B}(I^x(act'_p, M_q)), \tag{4.30}$$

$$s.t.\ \sum_{p} Cost(x_p, act'_p) \leq B_g, \tag{4.31}$$

$$\{x|x_p, p \in \{1, 2, ..., N\}\}, \tag{4.32}$$

$$\{M|M_q \in M, q \in \{1, 2, ..., B\}\}. \tag{4.33}$$

Chapter 5

APPLICATIONS

In this Chapter, the case studies considered within the Hierarchical Decision Making Framework for vulnerability to intentional hazards of the safety-critical systems are briefly illustrated. A relative evaluation of the vulnerability by ranking method is first shown (Section 5.1.1); An absolute evaluation of the vulerability is then given by applying the classification method with the corresponding sensitivity analysis (presented by confidence and accuracy of the assignment results) (Section 5.1.2); The inverse classification problem aiming to choose the set of "optimal" set of protective actions is tackled in two ways based on sensitivity indicators and optimization (Section 5.2); And finally, the inconsistency study of a giving set of data is demonstrated (Section 5.3). For further details the interested reader is referred to the corresponding Papers (i)-(v) of Part II.

5.1 Analysis of the vulnerability of a fleet of Nuclear Power Plants

Based on the methods presented in the previous Chapters, in this section, the vulnerability to intentional hazards is studied within a preliminary ranking framework, and then mainly the classification framework. The classification model is built to assign each alternative an pre-defined vulnerability class/category and the uncertainty of

the model is also assessed by giving an accuracy and confidence of the results.
We assume that the attributes in the data are categorical. In fact, for the basic layer criteria (presented in Chapter 2, detailed in Paper (i) and (ii)), there are several cases when the records in the data are quantitative. This is quite simple to be tackled with, since quantitative variables can be transformed to categorical form using discretization). We also note that since this technique will be used for action-driven applications for the amelioration part in which the features define the actions that may or may not happen, it is more natural to use categorical variables.

5.1.1 Vulnerability analysis by ranking

For illustration purposes, 9 fictitious plants (named F_1 to F_9) are considered to obtain the value functions, which are in turn used to evaluate the susceptibility to intentional attacks of 9 real plants (named R_1 to R_9). In simple words, the former 9 fictitious plants are evaluated with respect to their susceptibility to intentional hazards, to build the base for comparison of the latter. Best (least vulnerable) and worst (most vulnerable) fictitious plants are defined as bounding references, by taking the best/worst conditions of all subcriteria considered. The details are presented in Paper (i).
The analysis using ACUTA method needs a ranking of the fictitious NPPs to begin with. It is usually given by the experts. In our case study, the utility functions are first given by the authors. Let F be the set of the 9 fictitious plants and $F_L = \{F_j, j = 1, ..., 9\}$ the learning set. The data of $fictitiousWorst$ and $fictitiousBest$ are used to be the limit interval for the given criterion, divided into 5 subintervals. The utility functions are given such that all the data of $fictitiousWorst$ are set to 0 and the data sum of the $fictitiousBest$ is set to 1.
Based on the utility functions of the main criteria and the data, we can obtain the marginal value of the corresponding criterion for each fictitious NPP.
As a characteristic of the additive model, the global values which represent the susceptibility of the NPPs to intentional hazards are given by the sum of its marginal values. These values are used to rank the NPPs. The ranking obtained is integrated

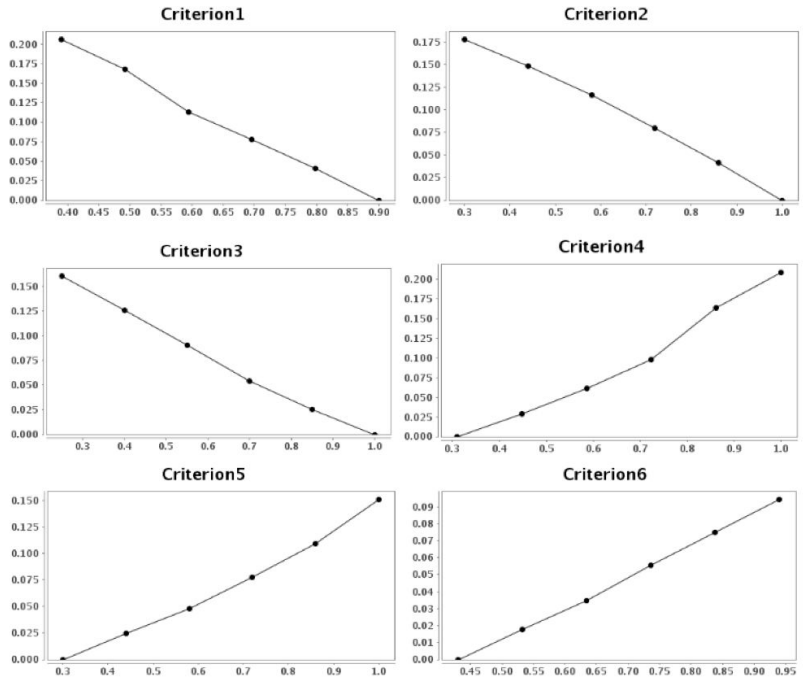

Figure 5-1: Representation of the Value Functions

into the decision-making process to find out the value functions for the 6 criteria through the ACUTA method. The intentional hazards of real plants is then analysed and represented by using Diviz (which is a software for designing, executing and sharing Multicriteria Decision Aid (MCDA) methods, algorithms and experiments. Based on basic algorithmic components, it allows combining these criteria for creating complex MCDA workflows and methods.). The value functions of the 6 main criteria (presented in Chapter 2) are shown in Figure 5-1.

The criteria preference directions can be recognised easily from the trends of the curves. Also, for most part of each curve, it is natural that the vertical axis values are roughly proportional to the abscissa axis ones. More importantly, we can figure out the sensitive interval of each criterion. This can be an indicator to know better the preference of the DMs during the ranking step and can also serve as a guidance during the amelioration of the plants.

In using the value functions, the former data of the 9 real NPPs can then be taken into account. We can compare the NPPs by single criterion. As shown in the 6 histograms (Figure 5-2), for one criterion, each column represents the corresponding performance

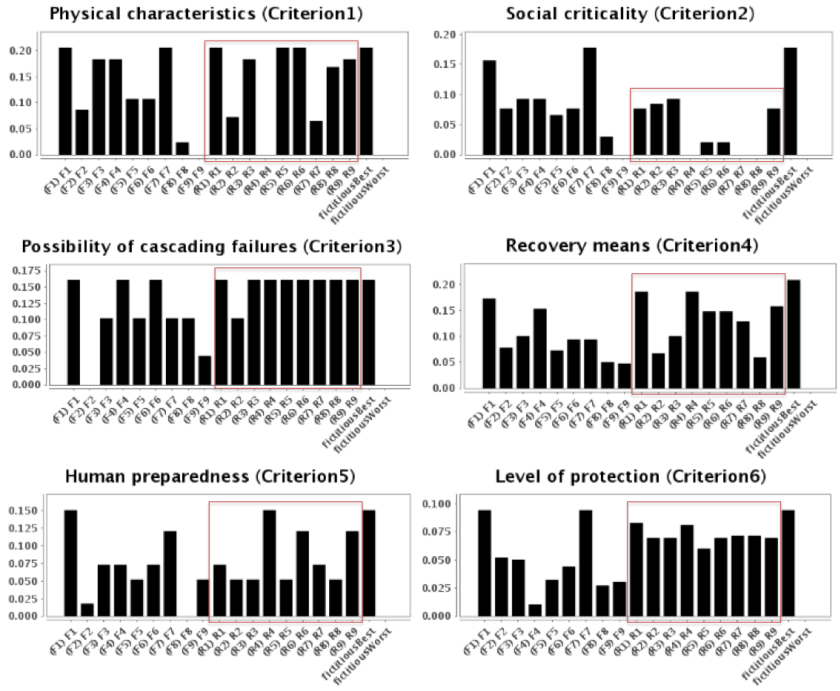

Figure 5-2: Histograms of subcriteria of the NPPs

of a given NPP. The length of each column is proportional to the marginal values. The longer the column, the better performance it has for the criterion. In the solid line frame there are the representative columns for each criterion of the real plants. As a characteristic of the additive model, the global values which represent the susceptibility of the NPPs to intentional hazards are given by the sum of the marginal values. An overview of the 20 NPPs is presented graphically in Figure 5-3. Each column represents the susceptibility performance of one NPP to intentional hazards. Each column is constituted by 6 blocks with different textures that represent the 6 main criteria. As mentioned before, the height of each block of the representative column is proportional to the value of the corresponding criterion data. The smaller the height of the representative column of a plant is, the more susceptible it is in facing an intentional hazard.

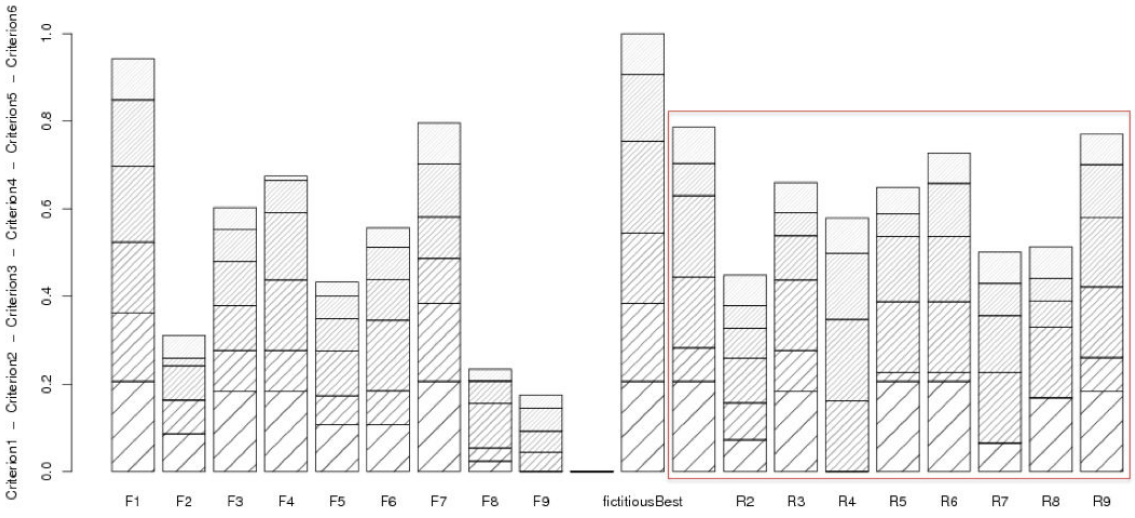

Figure 5-3: Histogram of susceptibility to intentional hazards of the NPPs

5.1.2 Vulnerability assessment by empirical classification

The methods presented in Chapter 3 & 4 are here applied on an exemplificative case study concerning the vulnerability analysis of Nuclear Power Plants (NPPs) [126]. We identify $n = 6$ main criteria $i = 1, 2, ..., n = 6$ by means of the hierarchical approach presented in [126], see Chapter 3: $MCrit_1$ = physical characteristics, $MCrit_2$ = social criticality, $MCrit_3$ = possibility of cascading failures, $MCrit_4$ = recovery means, $MCrit_5$ = human preparedness and $MCrit_6$ = level of protection. Then, $M = 4$ vulnerability categories $A^h, h = 1, 2, ..., M = 4$ are defined as: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic and A^4 = serious. The training set D_{TR} is constituted by a group of $N = 11$ NPPs, x_p, with the corresponding a priori-known categories Γ_p^t, i.e., $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N = 11\}$. The training set is summarised in Table 5.1.

In what follows, the three techniques of Chapter 3 are applied to assess the performance of the MR-Sort classification-based vulnerability analysis model built using the training set D_{TR} of Table 5.1.

Table 5.1: Training set with $N = 11$ assigned alternatives

Alternatives, x_p	Vulnerability Class Γ_p^t
$x_1 = \{0.61, 0.6, 0.75, 0.86, 1, 0.94\}$	A^1
$x_2 = \{0.33, 0.27, 0, 0.575, 0.4, 0.72\}$	A^3
$x_3 = \{0.55, 0.33, 0.5, 0.725, 0.7, 0.71\}$	A^2
$x_4 = \{0.55, 0.33, 0.75, 0.8, 0.7, 0.49\}$	A^3
$x_5 = \{0.39, 0.23, 0.5, 0.6, 0.6, 0.62\}$	A^3
$x_6 = \{0.39, 0.27, 0.75, 0.725, 0.7, 0.68\}$	A^2
$x_7 = \{0.61, 0.7, 0.5, 0.725, 0.9, 0.94\}$	A^2
$x_8 = \{0.16, 0.1, 0.5, 0.475, 0.3, 0.59\}$	A^4
$x_9 = \{0.1, 0, 0.25, 0.5, 0.6, 0.61\}$	A^4
$x_{10} = \{0.1, 0, 0, 0.3, 0.3, 0.43\}$	A^4
$x_{11} = \{0.61, 0.7, 0.75, 1, 1, 0.94\}$	A^1

Application of the Model Retrieval-Based Approach

We generate $N_{sets} = 1000$ different training sets $D_{TR}^{rand,j}, j = 1, 2, ..., N_{sets}$, and for each set j, we randomly generate $N_{models} = 100$ models $M(\cdot|\omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models} = 100$. By so doing, the expected accuracy (1-ϵ) of the corresponding MR-Sort model is obtained as the average of $N_{sets} \cdot N_{models} = 1000 \cdot 100 = 100000$ values $(1 - \epsilon_{jl}), j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$ (see Chapter 3.3.1). The size N_{test} of the random test set D_{TR}^{rand} is $N_{test} = 10000$. Finally, we perform the procedure of Chapter 3.3.1 for different sizes N of the random training set D_{TR}^{rand} (even if the size of the real training set available is $N = 11$, see Table 5.1): in particular, we choose $N = 5, 11, 20, 50, 100$ *and* 200. This analysis serves the purpose of outlining the behaviour of the accuracy $(1 - \epsilon)$ as a function of the amount of classification examples available.

The results are summarised in Figure 5-4 where the average percentage assignment error ϵ is shown as a function of the size N of the learning set (from 5 to 200). As expected, the assignment error ϵ tends to decrease when the size of the learning set N increases: the higher the cardinality of the learning set, the higher (resp. lower) the accuracy (resp. the expected error) in the corresponding assignments. It can be seen that for our model with $n = 6$ criteria and $M = 4$ categories, in order to guarantee an error rate inferior to 10% we would need training sets consisting of more than $N = 100$ alternatives. Typically, for a learning set of $N = 11$ alternatives (like that

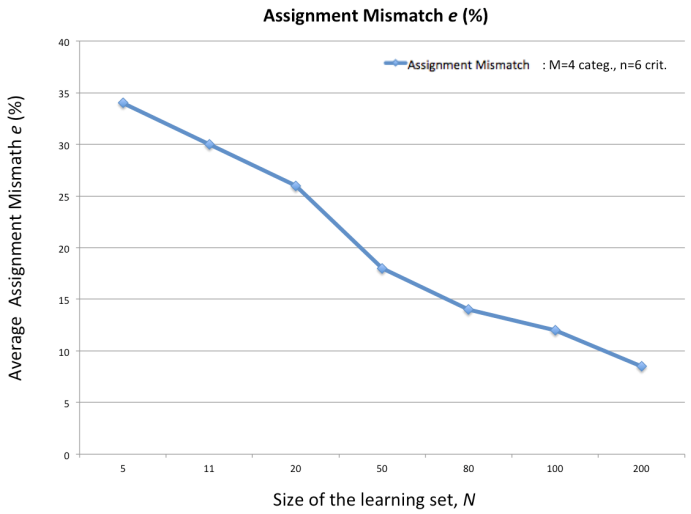

Figure 5-4: Average Assignment error ϵ (%) as a function of the size N of the learning set according to the model retrieval-based approach of Chapter 3.3.1

available in the present case study), the average assignment error ϵ is around 30%; correspondingly, the accuracy of the MR-Sort classification model trained with the data set D_{TR} of size $N = 11$ available in the present case is around $(1 - \epsilon) = 70\%$: in other words, there is a probability of 70% that a new alternative (i.e., a new NPP) is assigned to the correct category of vulnerability.

In order to assess the randomness intrinsic in the procedure used to obtain the accuracy estimate above, we have also calculated the 95% confidence intervals for the average assignment error ϵ of the models trained with $N = 11, 20$ *and* 100 alternatives in the training set. The 95% confidence interval for the error associated to the models trained with 11, 20 and 100 alternatives as learning set are [25.4%, 33%], [22.2%, 29.3%] and [10%, 15.5%], respectively. For illustration purposes, Figure 5-5 shows the distribution of the assignment mismatch built using the $N_{sets} \cdot N_{models} = 100000$ values $\epsilon_{jl}, j = 1, 2, ..., N_{sets} = 1000, l = 1, 2, ..., N_{models} = 100$, generated as described in Chapter 3.3.1 for the example of 11 alternatives.

Application of The Bootstrap Method

A number B (= 1000) of bootstrapped training sets $D_{TR,q}, q = 1, 2, ..., 1000$ of size $N = 11$ is built by random sampling with replacement from D_{TR}. The sets $D_{TR,q}$ are

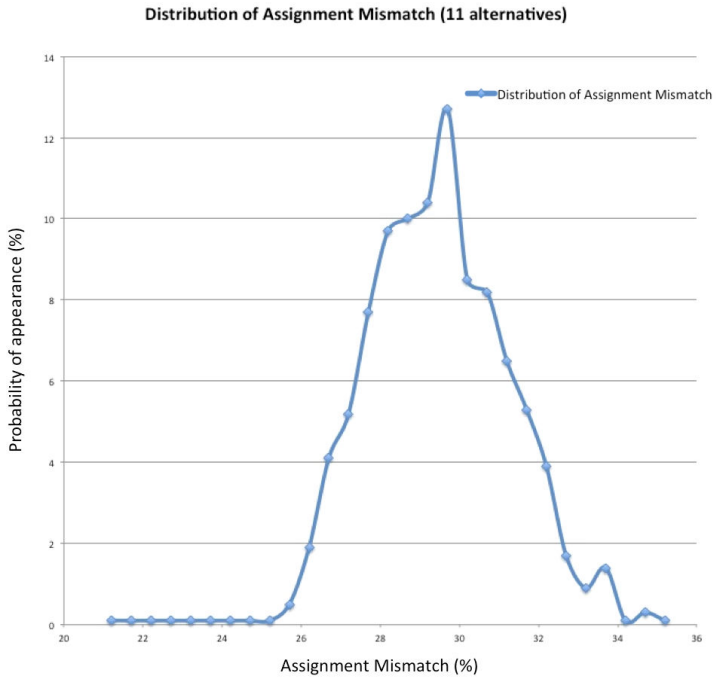

Figure 5-5: Distribution of the assignment mismatch for a MR-Sort model trained with $N = 11$ alternatives (%)

then used to train $B = 1000$ different classification models $\{M_1, M_2, ..., M_{1000}\}$. This ensemble of models can be used to classify new alternatives. Figure 5-6 shows the probability distributions $P(A^h|x_p), h = 1, 2, ..., M = 4, p = 1, 2, ..., N = 11$, empirically generated by the ensemble of $B = 1000$ bootstrapped MR-Sort classification models in the task of classifying the $N = 11$ alternatives of the training set $D_{TR} = \{x_1, x_2, ..., x_N\}$. The categories highlighted by the rectangles are those selected by the majority of the classifiers of the ensemble (Figure 5-6): it can be seen that the assigned classes coincide with the original categories of the alternatives of the training set (Table 5.1), i.e., the accuracy of the inferred classification model based on the given training set (with 11 assigned alternatives) is 1.

In order to investigate the confidence of the algorithm in the classification of the test patterns, the results achieved testing one specific pattern taken in turn from the training set are analysed. For each test of a specific pattern x_p, the distribution of the assignments by the $B = 1000$ classifiers shows the confidence of the assignment of the classification model on this specific pattern. By way of example, it can be

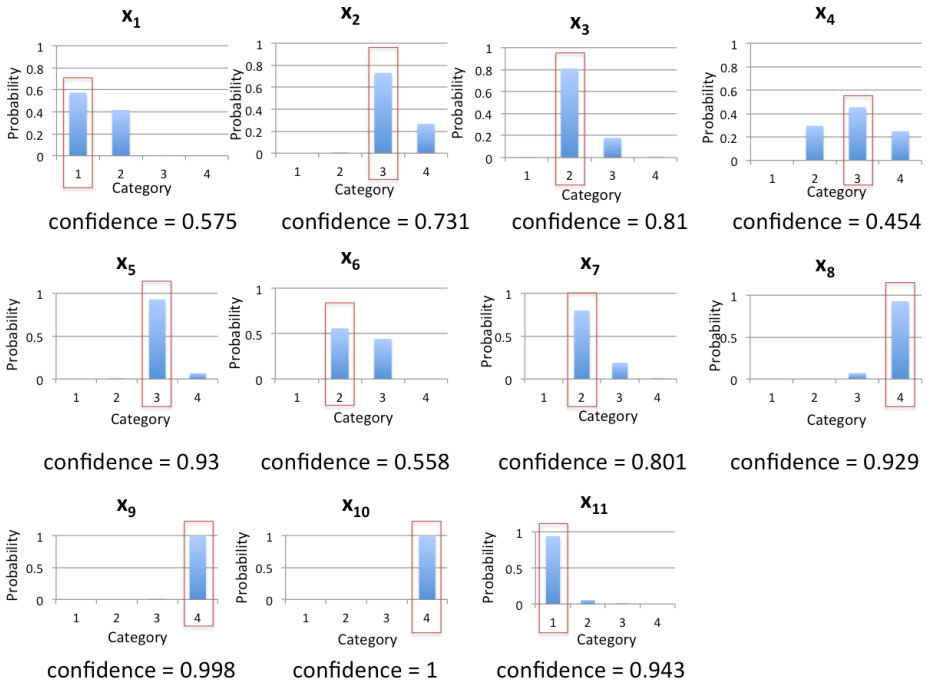

Figure 5-6: Probability distributions $P(A^h|x_p), h = 1, 2, ..., M = 4, p = 1, 2, ..., N = 11$ obtained by the ensemble of $B = 1000$ bootstrapped MR-Sort models in the classification of the alternatives x_p contained in the training set D_{TR}.

Table 5.2: Number of patterns classified with confidence value

Confidence range	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]
Number of patterns	1	2	0
Confidence range	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]
Number of patterns	1	2	5

seen that alternative x_3 is assigned to Class A^2(the correct one) with a confidence of $P(A^2|x_3) = 0.81$, whereas alternative x_6 is assigned to the same class A^2, but with a confidence of only $P(A^2|x_6) = 0.56$.

Notice that the most interesting information regards the confidence in the assignment of the test pattern to the class with the highest number of votes, i.e., the class actually assigned by the ensemble system according to the majority voting rule adopted [10]. In this respect, Table 5.2 reports the distribution of the confidence values associated to the class to which each of the 11 alternatives has been assigned.

Thus, a $10/11 \approx 91\%$ of all class assignments with confidence bigger than 0.5 are correct.

Table 5.3: Comparison between the real categories and the assignments provided by the LOOCV models

Alternative	Real Categories, Γ_p^t	Assignments by LOOCV method
x_1	1	1
x_2	3	3
x_3	2	2
x_4	3	2
x_5	3	3
x_6	2	3
x_7	2	2
x_8	4	4
x_9	4	4
x_{10}	4	4
x_{11}	1	1

Application of the Leave-One-Out Cross-Validation (LOOCV) Method

Based on the original training set D_{TR} of size $N = 11$, we generate 11 "new" training sets $D_{TR,i}, i = 1, 2, ..., 11$ (each containing $N - 1 = 10$ assigned alternatives) by taking out each time one of the alternatives from D_{TR}. These 11 training sets are, then, used to train 11 different classification models $M_1, M_2, ..., M_{11}$. Each of these 11 models is used to classify the alternative correspondingly taken-out. Table 5.3 shows the comparison between the real classes Γ_p^t of the alternatives of the training set and the categories assigned by the trained models. It can be seen that $\epsilon = 2$ out of the $N = 11$ alternatives are assigned incorrectly (alternatives x_4 and x_6). Thus, the accuracy in the classification is given by the complement to 1 of the average error rate, i.e., $1 - \epsilon/N = 1 - 2/11 = 1 - 0.182 = 0.818$. Notice that the 95% confidence interval for this recognition rate is [0.5901, 1].

For more details, the interested readers are invited to consult Paper (ii) in Part II.

Table 5.4: Available protective actions

No.	Direct action effect
act^1	Reduce number of workers
act^2	Decrease nominal power production
act^3	Cut down number of production units
act^4	Decrease percentage of contribution to the welfare
act^5	Add installed backup components
act^6	Add external emergency measures
act^7	Prolong the duration of backup components
act^8	Reduce the duration of repair and recovery actions
act^9	Enhance training
act^{10}	Strengthen safety management
act^{11}	Decrease number of accesses
act^{12}	Enhance entrance control
act^{13}	Strengthen surveillance

5.2 Identification of protective actions to reduce the overall vulnerability of a fleet of Nuclear Power Plants

The inverse classification model serves to generate an "optimal" set of protective actions for the considered group of Nuclear Power Plants (NPPs).

5.2.1 Choice of the set of protective actions by means of sensitivity indicators

As presented in Chapter 4, the choice of the set of protective actions under a budget limitation is structured as an inverse classification problem. As shown in Figure 4-1 presented in Chapter 4, we define $F = 13$ direct actions ($act = \{act^1, act^2, ..., act^k, ..., act^F\}$), each acting on one or more subcriteria (Table 5.4). The influence of each of the actions have multiple influences on different criteria, with possibly positive or negative effects and quantified by the different weights/coefficients. Also, for each action we consider different levels of implementation l_p^k ($l_p^k, l_p^k \in \{0, 1, 2, 3\}, p \in \{1, 2, ..., N\}, k \in \{1, 2, ..., F\}$), representing to what extent/how far/in which amount action k is applied on system p (notice that $l_p^k = 0$ means that action k is not applied to system p).

Finally, for simplicity we assume that the cost related to the application of a given action is proportional to the level of the action ($c_k^p(l_p^k) = l_p^k, k \in \{1, 2, ..., F\}, p \in \{1, 2, ..., N\}$). More details of the action related parameters can be found in appended Papers (iv) and (v).

In what follows, an analyses (based on the sensitivity indicators presented in Chapter 4) of different combinations of actions are ranked according to their ability in reducing the vulnerability of a group of NPPs.

Ranking different combinations of actions based on $\overline{\Delta M}$

A set x of $N(N = 20)$ NPPs ($x = \{x_p, p \in \{1, 2, ..., N\}\}$) is available: 10 of them (NPPs from x_6 to x_{15}) are selected as a reference set x^{ref} to evaluate the sensitivity indicators; the remaining NPPs are regrouped to form a set $x^{test}(x^{test} = \{x_p, p \in \{1, 2, ..., 5\} \cup \{16, 17, ..., N\}\})$ used to test the combinations of actions ranked using x^{ref}. Based on the reference set, we have performed an exhaustive calculation of the value of $\overline{\Delta M}$ for all the possible combinations of actions (in total, 4^{13} combinations). Then, we selected the ones (in total 29940 combinations) that have the (same) highest value of $\overline{\Delta M}$ (i.e., $\overline{\Delta M} = 14$): these represent the optimal combinations of actions according to $\overline{\Delta M}$: in what follows, this set is referred to as $act_{\overline{\Delta M}}$.

All the combinations of actions belonging to the set $act_{\overline{\Delta M}}$ are applied to each of the $N(N = 20)$ NPPs in x: the resulting categories ($A_p^{\lambda'}, p \in \{1, 2, ..., N\}$) are reported in Table 5.5. Note that the actions are ranked according to values of $\overline{\Delta M}$ that are evaluated on a *group* of *reference plants* (x^{ref}): in this view, they provide an indication only on the *expected* performance of the actions on *new* plants and, thus, they may not provide any indications about the combination of actions that is *optimal* for *one particular* plant. Thus, in order to verify how close these sets of actions are to the combinations that are optimal for a particular NPP, we compare the assignments $A_p^{\lambda'}$(Table 5.5) with the best category that each NPP may reach ($A_p^{\lambda}, p \in \{1, 2, ..., N\}$) (in other words, A_p^{λ} is the category that x_p reaches after the application of a combination of actions that is *the optimal one* for *that particular* plant). In order to do so, another exhaustive calculation is done upon the group x with the purpose of finding

the actions that bring each particular NPP to the best category possible (notice that for some NPPs, reaching A^1 may not be possible). All the possible combinations of actions are tested on each NPP in order to find the best assignment A_p^λ for each of them. The results are shown in Table 5.5. The first column of the results shows the original assignments for the NPPs in the studied set x. The second column shows the corresponding possibly best assignments A^λ and the third column provides the new assignments $A^{\lambda'}$ after the application of the combinations of actions included in $act_{\overline{\Delta M}}$.

Analyzing the best assignments A^λ of the NPPs in the reference set $A_p^\lambda, p \in \{6, 7, ..., 15\}$, we observe that they coincide perfectly (100%) with the assignments $A^{\lambda'}(A_p^{\lambda'}, p \in \{6, 7, ..., 15\})$ obtained after the application of the actions in $act_{\overline{\Delta M}}$. If we take the NPPs in the test set as new NPPs and compare the assignments obtained by these two methods with the original assignments $A(A_p, p \in \{1, 2, ..., 5\} \cup \{16, 17, ..., N\})$, we find that: (i) all the NPPs are stable or ameliorated after the application of the combinations of actions in $act_{\overline{\Delta M}}$; (ii) there are 2 out of 10 NPPs that are not ameliorated to the best category A_p^λ (i.e., x_{16} *and* x_{19}): they remain in the same category; instead, 8 out of 10 NPPs are ameliorated to their best possible categories: then, the probability that the combinations of actions $act_{\overline{\Delta M}}$ ameliorate a new NPP to its best possible category A^λ is 80%.

The analysis of inverse classification problem tackled using the sensitivity indicators and taking into account the action costs and budget limitations is not presented in this thesis work. For further details the interested reader is referred to the corresponding Paper (iv) of Part II.

5.2.2 Choice of the set of protective actions with a limited budget by optimization

After the analysis of the previous section we can have a rough idea of the capacity of amelioration all combination of actions can have upon a group of alternatives. In

Table 5.5: Comparison of assignments: Best possible Assignment A_p^{λ} and After-action Assignment $A_p^{\lambda'}$ listed with NPPs that are differently assigned highlighted (x_{16}, x_{19}).

No.	Original Assignment	Best possible Assignment A_p^{λ}	After action Assignment $A_p^{\lambda'}$
x1	1	1	1
x2	3	3	3
x3	2	2	2
x4	3	1	1
x5	3	2	2
x6	2	1	1
x7	2	1	1
x8	4	2	2
x9	4	2	2
x10	4	3	3
x11	1	1	1
x12	2	1	1
x13	3	2	2
x14	3	1	1
x15	4	1	1
x16	3	2	3
x17	2	2	2
x18	3	2	2
x19	3	2	3
x20	2	1	1

what follows, carrying on the characteristics of the model, the inverse classification problem is handled by the optimization approach with different perspectives (simple optimization, robust optimization and probabilistic optimization) in order to choose the set of protective actions to minimize the overall level of vulnerability of a group of safety-critical systems with budget constraints. The training set D_{TR} in this section is constituted by a group of $N = 18$ NPPs with the corresponding a priori-known categories Γ_p^t, i.e., $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N = 18\}$. (We use the same training set as in the previous section but considering only the NPPs that are pre-assigned to A^2, A^3 *or* A^4, since the alternatives originally assigned to the best category A^1 are not taken into account. The sequence numbers of the alternatives are also rearranged.)

Simple Optimization

Two tests are first carried out considering an unlimited or limited budget. The example with unlimited budget aims at showing an ideal case of the inverse classification problem that would lead, in principle, to the best after-action condition (same idea as the exhaustive calculation in the previous section and the correponding best possible after-action categories for all alternatives are also coherent.). It can be seen that, based on the original dataset of pre-assigned alternatives, there are certain NPPs (i.e.,

Table 5.6: After-action assignments of the considered NPPs without budget constraint. White cases in the third column indicate unchanged assignment.

Unlimited budget	Original assignment	Best after-action assignment
x1	3	3
x2	2	2
x3	3	1
x4	3	2
x5	2	1
x6	2	1
x7	4	2
x8	4	2
x9	4	3
x10	2	1
x11	3	2
x12	3	1
x13	4	1
x14	3	2
x15	2	2
x16	3	2
x17	3	2
x18	2	1

x_1, x_2 *and* x_{15}) that can never be ameliorated to the best category A^1 (Table 5.6). The identification of the best (i.e., lowest) vulnerability category that one NPP can be assigned to without budget restrictions represents an important base information that provides the decision makers with a global view of the problem goals.

The optimization performed with budget constraints aims at solving the realistic problem of finding out the combination of protective actions for each NPP, that ameliorate the groupe of NPPs with priority to the most vulnerable ones, managing the "residual" resource to improve the others. With an unlimited budget, most of the NPPs are ameliorated to a lower level of vulnerability. Actually, x_1, x_2 *and* x_{15} do not change class because of their particular characteristics (e.g., the physical distance between the site and the nearby cities is closer with respect to that of the other plants, and such characteristic cannot be modified by any action). The minimum cost necessary to improve each NPP to the best possible category is $B_{g_{min}} = 78$. Fixing a limited budget to $B_g = 40$, the optimization of the actions leads to the ameliorations reported in Table 5.7. Obviously, x_1, x_2 *and* x_{15} still do not change class as in the

Table 5.7: After-action assignments of the considered NPPs with budget constraint $B_g = 40$(simple optimization). White cases in the third column indicate unchanged assignment.

Budget constraint: Bg=40	Original assignment	Best after-action assignment
x1	3	3
x2	2	2
x3	3	1
x4	3	2
x5	2	2
x6	2	1
x7	4	2
x8	4	3
x9	4	3
x10	2	2
x11	3	2
x12	3	2
x13	4	2
x14	3	2
x15	2	2
x16	3	2
x17	3	3
x18	2	1

case with unlimited budget. Moreover, since the budget is lower than that necessary to ameliorate all NPPs to their best category ($B_{g_{min}} = 78$), there are other NPPs (x_5, x_{10} *and* x_{17}) whose vulnerability category is not changed. On the contrary, all NPPs originally assigned to the "worst" category A^4 improve after action(s); then, the rest of the budget is distributed to ameliorate the other NPPs as much as possible with the given budget. For example, x_8 *and* x_{12} are improved by one category, whereas they can be improved by two categories in the case of unlimited budget (Table 5.6).

In the next two subsections, we present the results of the other two optimization approches, considering only the realistic case of limited budget.

Robust Optimization

The results in the case of limited budget, $B_g = 40$ are shown in Table 5.8 and compared to the original categories (obtained by majority-voting over the B compatible

Table 5.8: After-action assignments of the considered NPPs with budget constraint (robust optimization). White cases in the third column indicate unchanged assignment. MV = majority-voting.

Budget constraint: Bg=40	Original assignment by MV	Robust optimization after-action by MV
x1	3	3
x2	3	2
x3	3	2
x4	3	3
x5	3	3
x6	2	2
x7	4	4
x8	4	4
x9	4	4
x10	2	2
x11	3	3
x12	3	3
x13	4	3
x14	3	3
x15	4	4
x16	3	3
x17	4	4
x18	3	2

bootstrapped classification models). There are only 4 NPPs that are ameliorated: x_{13} is ameliorated from A^4 to A^3; x_2, x_3 *and* x_{18} are ameliorated from A^3 to A^2. There are changes in the bootstrapped distributions of the categories of the other NPPs, but not consistent enough to change their final assignments by majority-voting. In comparison with the results obtained in the previous subsection, there are less NPPs that are ameliorated. This is reasonable for a "robust" solution, since "extreme" (worst-case) compatible classification models affect the optimization.

Probabilistic Optimization

The probabilistic optimization is a variation of the Robust optimization. Instead of maximizing $Min_M(I^x(act'_p, M_q))$ (i.e., the worst after-action objective function value), we choose to maximize the expected value of the bootstrapped probability distribution of the weighted objective function $I^x(act'_p, M_q)$.

The results are shown in Table 5.9, in comparison with the original majority-voting

Table 5.9: After-action assignments of the considered NPPs with budget constraint (probabilistic optimization). White cases in the third column indicate unchanged assignment. MV = majority-voting.

Budget constraint: Bg=40	Original assignment by MV	Probabilistic optimization after-action by MV
x1	3	3
x2	3	2
x3	3	2
x4	3	3
x5	3	2
x6	2	2
x7	4	4
x8	4	3
x9	4	4
x10	2	2
x11	3	3
x12	3	2
x13	4	3
x14	3	3
x15	4	3
x16	3	3
x17	4	3
x18	3	3

category of each NPP. There are 8 NPPs that are ameliorated: x_8, x_{13}, x_{15} *and* x_{17} are changed from category A^4 to A^3; x_2, x_3 *and* x_{12} are changed from A^3 to A^2. In comparison with the results of simple optimization, there are less NPPs that are ameliorated; in addition, not all the NPPs that were originally assigned to the worst category (A^4) are improved. On the other hand, with respect to the results of the robust optimization (which also considers an ensemble of different compatible models), the group of NPPs is globally improved. The results of the probabilistic case are more satisfactory since most of the NPPs that were assigned to the worst category (A^4) are improved; then, the rest of the resources is used to ameliorate those plants that were assigned to the second worst category (A^3).

More details of the three optimization approaches can be found in the appended Paper (v) in Part II.

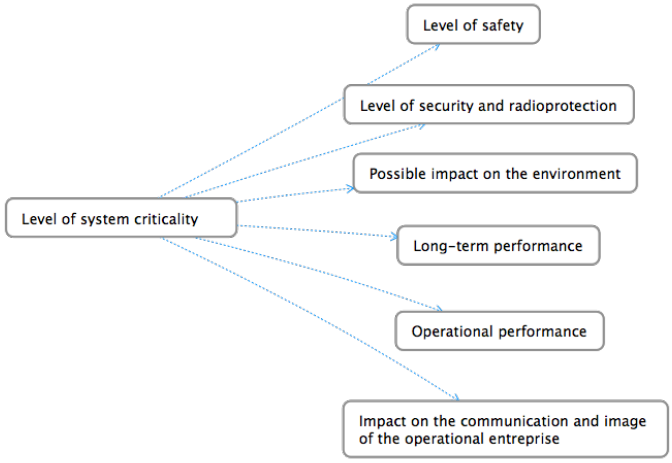

Figure 5-7: Criteria used to characterize the overall level of criticality of a complex energy production system or plant.

5.3 Analysis of data inconsistency

In this case study, based on the characteristic of the given data (represented directly by 6 main criteria), the hierarchical model is simplified and adapted to the problem, we consider that the overall level of criticality of the system is characterized in terms of: $MCrit_1$ = level of safety, $MCrit_2$ = level of security and protection, $MCrit_3$ = possible impact on the environment, $MCrit_4$ = long-term performance, $MCrit_5$ = operational performance and $MCrit_6$ = possible impact on the communication and image of the operational enterprise, as shown in Figure 5-7: each criterion is evaluated in 4 grades, ranging from best (grade '0') to worst (grade '3'). Four levels (or categories) of criticality are considered: satisfactory (0), acceptable (1), problematic (2) and serious (3). Then, the assessment of the level of criticality can be performed within a classification framework as in the previous case study. More details including the "scoring" of the criticality of each criterion are presented in Paper (iii).

The application of the MR-sort disaggregation algorithm on the given set of alternatives (Table 5.10)[1] does not lead to the generation of any classification model, because there are inconsistencies within the given data. There may exist different types of

[1]Data provided by EDF R&D - Management des Risque industriels.

	Criticality evaluation criteria						
Alternatives (NPPs)	Safety	Secrurity and Radioprotection	Impact on the Environment	Long-term Performance	Operational Performance	Communication and Image of the Operational Enterprise	Category Assignment (original set)
x1	3	0	3	3	0	2	3
x2	1	0	1	1	0	2	1
x3	1	0	1	2	0	1	2
x4	2	2	3	0	0	1	2
x5	3	1	2	3	0	1	3
x6	1	3	2	2	0	1	2
x7	2	0	3	2	0	3	4
x8	2	2	3	2	0	0	1
x9	1	0	2	0	0	0	1
x10	2	0	3	0	0	2	3
x11	2	0	3	2	0	2	3
x12	1	0	3	1	0	1	3
x13	1	0	2	0	0	1	1
x14	2	0	0	0	0	1	2
x15	1	0	0	0	0	0	1
x16	1	0	0	0	0	1	3
x17	2	0	0	2	0	1	3
x18	1	2	2	0	0	1	2
x19	0	0	0	0	0	1	3
x20	0	3	0	0	1	0	4
x21	1	0	2	1	1	0	4
x22	1	3	0	0	1	0	2
x23	1	0	1	0	1	0	1
x24	1	0	2	0	0	0	4
x25	1	0	0	0	1	0	1
x26	1	0	0	0	0	0	2
x27	1	0	0	0	0	1	2
x28	1	0	0	0	0	1	2
x29	2	2	3	0	0	0	3
x30	2	2	3	2	0	0	2
x31	2	2	2	1	0	0	1
x32	3	0	3	0	0	3	2
x33	1	0	1	0	0	0	3
x34	3	0	0	1	0	3	3
x35	3	0	0	0	0	3	2

Table 5.10: Original training data set

inconsistencies, e.g., two alternatives (x_{16} and x_{27}) with same value for all the six criteria are assigned to different categories (resp., 3 and 2), an alternative (x_{19}) with better characteristics than another (x_{13}) with respect to the six criteria, is assigned to a worse category (3), etc.

Such inconsistencies are solved via constraints deletion (subsection 5.3.1) and constraints relaxation (subsection 5.3.2).

5.3.1 Inconsistency resolution via constraints deletion

We first consider finding out the consistent dataset with maximized number of pre-assigned alternatives. We analyze the given data set by the constraints deletion algorithm. In the given set D of 35 alternatives, 14 are deleted, which leaves a consistent data set of 21 alternatives. The new consistent set $D_{ad} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{ad} = 21\}$ is, then, used to generate a compatible classification model by the

MR-sort disaggregation algorithm. Then, all the alternatives in the original data set D are assigned a class by model M_{ad}: such assignments agree with the results of the constraints deletion process, i.e., only the deleted alternatives are not correctly assigned (see Table 5.11, where the deleted alternatives are highlighted).

5.3.2 Inconsistency resolution via constraints relaxation

In the previous Section, we succeeded in obtaining a consistent data set from a given inconsistent one by deleting the inconsistent alternatives of a "wrong" assignment. However, from the point of view of the decision makers, it would be ideal to retain as many alternatives as possible in the training set, especially when the size of the ensemble is limited (which is always the case of the evaluation problem of safety-critical infrastructures). This can be done by modifying the pre-defined (wrong) assignments of the inconsistent alternatives.

We examine the same set D by means of the constraints relaxation algorithm presented in Chapter 3.2.3. After the application of the algorithm, we obtain the set $D_{ar} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{ar} = 21\}$, which is identical to the set $D_{ad} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{ad} = 21\}$ obtained in the previous subsection (for the alternatives in this set, the corresponding generated constraints are consistent). The remaining alternatives form the set $D_r = D - D_{ar}$. However, this algorithm also allows the identification of two more sets: (i)$D_{up} = \{(x_p, \Gamma_p^t)|\gamma_p^+ = 0\}$ (i.e., the set of alternatives whose assignments should be better than the original one, indicated in Table 5.11 by a "+" in the shadowed Table cells in column "Constraint relaxation"); (ii)$D_{down} = \{(x_p, \Gamma_p^t)|\gamma^- = 0\}$(i.e., the set of alternatives whose assignments should be worse than the original one, indicated in Table 5.11 by a "-" in the shadowed Table cells in the column "Constraints relaxation").

Based on the indications given by the sets D_{up} and D_{down}, we have modified each of the alternatives in D_r by one category in the direction suggested by the relaxation algorithm. Combining the alternatives thereby modified in D_r with the ones in D_{ar}, we obtain a new data set of 35 alternatives $D_{relax} = \{(x_p, \Gamma_p^{relax} : p = 1, 2, ..., N_{relax} = 35)\}$, based on which a compatible classification can be generated.

	Criticality evaluation criteria								
Alternatives (NPPs)	Safety	Secrurity and Radioprotection	Impact on the Environment	Long-term Performance	Operational Performance	Communication and Image of the Operational Enterprise	Category Assignment (original set)	Constraints deletion	Constraints relaxation
x1 (TR)	3	0	3	3	0	2	3	3	3
x2 (TR)	1	0	1	1	0	2	1	2	-
x3 (TR)	1	0	1	2	0	1	2	2	2
x4 (TR)	2	2	3	0	0	1	2	3	-
x5 (TR)	3	1	2	3	0	1	3	3	3
x6 (TR)	1	3	2	2	0	1	2	2	2
x7 (TR)	2	0	3	2	0	3	4	4	4
x8 (TR)	2	2	3	2	0	0	1	3	-
x9 (TR)	1	0	2	0	0	0	1	1	1
x10 (TR)	2	0	3	0	0	2	3	3	3
x11 (TR)	2	0	3	2	0	2	3	3	3
x12 (TR)	1	0	3	1	0	1	3	3	3
x13 (TR)	1	0	2	0	0	1	1	2	-
x14 (TR)	2	0	0	0	0	1	2	2	2
x15 (TR)	1	0	0	0	0	0	1	1	1
x16 (TR)	1	0	0	0	0	1	3	2	+
x17 (TR)	2	0	0	2	0	1	3	3	3
x18 (TR)	1	2	2	0	0	1	2	2	2
x19 (TR)	0	0	0	0	0	1	3	1	+
x20 (TR)	0	3	0	0	1	0	4	1	+
x21 (TR)	1	0	2	1	1	0	4	1	+
x22 (TR)	1	3	0	0	1	0	2	2	2
x23 (TR)	1	0	1	0	1	0	1	1	1
x24 (TR)	1	0	2	0	0	0	4	1	+
x25 (TR)	1	0	0	0	1	0	1	1	1
x26 (TS)	1	0	0	0	0	0	2	1	+
x27 (TS)	1	0	0	0	0	1	2	2	2
x28 (TS)	1	0	0	0	0	1	2	2	2
x29 (TS)	2	2	3	0	0	0	3	3	3
x30 (TS)	2	2	3	2	0	0	2	3	-
x31 (TS)	2	2	2	1	0	0	1	1	1
x32 (TS)	3	0	3	0	0	3	2	3	-
x33 (TS)	1	0	1	0	0	0	3	1	+
x34 (TS)	3	0	0	1	0	3	3	3	3
x35 (TS)	3	0	0	0	0	3	2	3	-

Table 5.11: Original inconsistent dataset and the corresponding modifications operated by the constraint deletion and relaxation algorithms

In order to assess the performance of the classifications in terms of accuracy and confidence in the assignments based on the set which contains "relaxed" data, a group of $N = 25$ data of D_{relax} (marked as "TR" in the first column of Table 5.11) is used to build the training set D_{TR} for the model, i.e., $D_{TR} = \{(x_p, \Gamma_p^{relax} : p = 1, 2, ..., N = 25)\}$; the remaining 10 alternatives (marked as "TS" in the first column of Table 5.11) are used for testing the model generated. The three approaches used in the previous case study (namely, a model-retrieval-based approach, the bootstrap method, and the Leave-One-Out Cross-Validation (LOOCV) technique) are applied. For further details the interested reader is referred to the corresponding Paper (iii) of Part II.

Chapter 6

CONCLUSIONS AND FUTURE RESEARCH

6.1 Conclusions

This dissertation focuses on the vulnerability assessment and management of safety-critical complex systems, with respect to their vulnerability to intentional hazards (i.e., malevolent acts). A number of methods have been considered to represent, rank, assess and finally improve the performance of safety-critical systems. A comprehensive methodology has been developed, which combines: (i) the representation of the considered safety-critical complex systems within a hierarchical framework, where vulnerability is decomposed into the factors that influence it; (ii) the qualitative and quantitative evaluation of the vulnerability by two methods: (1) a ranking approach for a "comparative" evaluation of the vulnerabilities of different systems; (2) an empirical classification model to provide a quantitative "absolute" assessment of the level of vulnerability of each system; (iii) the assessment of the performance of the classification model in terms of accuracy and confidence in the assignments, in order to cope with uncertainties; (iv) the identification and resolution of possible inconsistencies in the data sets used to build the classification model; (v) the choice of protective actions for each system in order to minimize their vulnerability under a limited budget, by means of sensitivity indicators and optimization-based approaches.

A challenge related to the study of any complex system lies in the inherent complexity itself; thus, well-defined system boundaries and simplifications of the system representation and analysis are usually required. Based on recent developments in the field of complex system representation, this dissertation has introduced a method for the hierarchical "decomposition" and analysis of a (group of) safety-critical system(s) with respect to their vulnerability properties: this has been done within an "all-hazard" persepctive and with main focus on the susceptibility to intentional hazards. The availability of different scales of modeling resolution can be leveraged efficiently to facilitate the management of complexity in the analysis of large-scale complex systems. The applications to case studies involving nuclear power plants (NPPs) have demonstrated the effectiveness of the proposed method, in particular in identifying the most important contributions to vulnerability.

Two approaches (based on ranking and classification) have been introduced in order to provide a relative and an absolute evaluation of the vulnerability of a group of systems, respectively. On the one hand, a case study of NPPs has been analysed by using the Analytic Center UTilité Additive (ACUTA) method and the results calculated with the software Diviz. A ranking of vulnerability "levels" of a given group of plants has been, thus, obtained. On the other hand, a Majority Rule Sorting (MR-Sort) classification model has been trained by means of a small-sized set of data, representing a priori-known classification examples, to assign each alternative to a pre-defined vulnerability class. The performance of the MR-Sort classification model has been evaluated with respect to: (a) its classification accuracy (resp., error), that is, the expected fraction of systems correctly (resp., incorrectly) classified; (b) the confidence associated to the classification assignments (defined as the probability that the vulnerability class assigned by the model to a given (single) system is the correct one). The performance of the empirically constructed classification model has been assessed by resorting to three approaches: a model-retrieval-based approach, the bootstrap method, and the leave-one-out cross-validation (LOOCV) technique. To the best of the authors' knowledge, it has been the first time that:

- a classification-based hierarchical framework has been applied for the analysis

of the vulnerability of safety-critical systems to intentional hazards;

- the confidence in the assignments provided by an MR-Sort classification model has been quantitatively assessed by the bootstrap method, in terms of the probability that a given alternative is correctly classified.

From the results obtained, it has been concluded that although the model-retrieval-based approach may be useful for providing an upper bound on the error rate of the classification model (obtained by exploring the space of all possible compatible random models and training sets), the bootstrap method seems preferable for the following reasons: (i) it makes use of the training data set available from the particular case study at hand, thus characterizing the uncertainty intrinsic in it; (ii) for each alternative (i.e., safety-critical system) to be classified, it is able to assess the confidence in the classification by providing the probability that the selected vulnerability class is the correct one. This is of paramount importance in the decision-making processes involving the vulnerability assessment of safety-critical systems, since it provides a metric for quantifying the "robustness" of a given decision.

The problems of inconsistency and contradictions in the initial training data set have been addressed in order to tackle the "impossibility" to generate a compatible, "feasible" classification model by means of the MR-Sort method in the presence of incoherent data. Specifically, we have introduced a binary variable to describe whether one or more constraints provided by the data available should be considered or not in the generation of a compatible classification model. Two algorithms are developed to maximize the number of consistent data examples in the training set by deleting or relaxing, respectively, some constraints in the process of model construction.

Finally, a pragmatic inverse classification framework has been proposed in order to identify protective action(s) that reduce the vulnerability of safety-critical systems with respect to intentional hazards: the framework is based on the MR-Sort classification model presented before. Two different approaches of increasing complexity have been developed. Sensitivity indicators have been first introduced to evaluate and rank different combinations of actions with respect to their "expected" ability to

reduce the vulnerability of the safety-critical systems considered. A case study referring to NPPs vulnerability to intentional attacks has been worked out. The results have shown that the actions ranked as best according to the proposed indicators give a satisfactory performance in terms of reduction of vulnerability in test NPPs. Then, the issue has been tackled within an optimization framework. In particular, the set of protective actions to implement is chosen as the one minimizing the overall level of vulnerability of a group of safety-critical systems. In this context, three different optimization approaches have been explored: (i) one single classification model has been built to evaluate and minimize system vulnerability; (ii) an ensemble of compatible classification models, generated by the bootstrap method, has been employed to perform a "robust" optimization, taking as reference the "worst-case" scenario over the group of models; (iii) finally, a distribution of classification models, still obtained by bootstrap, has been considered to address vulnerability reduction in a "probabilistic" fashion (i.e., by minimizing the "expected" vulnerability of a fleet of systems). To the best of the authors' knowledge, it is the first time that an inverse classification approach has been applied for the optimal selection of the choice of protective actions to apply to each considered safety-critical systems (e.g., Nuclear Power Plants), considering different classification models and optimization approaches. From the results obtained, it can be concluded that all the proposed optimization algorithms work properly. Although the approach presented in the robust case may be useful for providing a "conservative" choice of actions ("regardless" of any particular compatible classification model), the formulation of the probabilistic case seems to be preferable for real cases for the following reasons: (i) as the robust case, it makes use of the training data set available from the particular case study at hand, thus characterizing the uncertainty intrinsic in it; (ii) with the objective of minimizing the "expected" level of vulnerability over all the systems considered, the "extreme" and worst-case scenarios (represented by "pathologic" models) are "neglected" or given "less weight", which is reasonable in real-world projects.

The proposed methodological framework provides a powerful tool for systematically and pragmatically evaluating the vulnerability of complex, safety-critical systems.

The study of the inverse classification problem is of paramount importance in the decision-making processes involving the choice of protective actions under different budgets constraints and the evaluation of the properties of the safety-critical systems after the application of such actions.

6.2 Future research

Some limitations and open problems arising from this dissertation necessitate discussion for possible further study. Firstly, the hierarchical representation of the vulnerability proposed in Chapter 2 is partial and is not to be considered exhaustive. Other properties such as the 'cyber characteristics' should also be taken into account to better describe the susceptibility to intentional hazards. Cyber security refers to the prevention and mitigation of the cyber threats beforehand and the appropriate response if a cyber-attack occurs. Nuclear facilities have serious concerns regarding cyber-attacks because of the vast and long-term effects of dangerous radioactive materials when an accident occurs [14][63][101][30][113][69].

In addition, the ranking model described in Chapter 2 gives a relative comparison of system vulnerability among the set of considered alternatives. An absolute evaluation by a classification model is preferred with respect to its practical significance. In this view, different classification models could be applied to different problems. In our case, the vulnerability classes are given as discrete precise numbers ($\{A^h : h = 1, 2, ..., M\}$). However, while for certain alternatives the desired after action classes are well defined (i.e., limited to one category), for other alternatives they may be more vague and imprecise (e.g., for one alternative, which is originally assigned to a good category, it may not be so important to define precisely a "target" category of amelioration). In such cases, fuzzy set theory [131][134][135][136][79] may be applied. A *membership (characteristic) function* $f_S^{A^h}(x)$ of a *fuzzy set* S can be generated to associate a "degree of membership" of the points $x \in S$ to class/category A^h. In the context of interest to the present thesis, x may represent an aggregated scalar value (obtained by Multi-Attibute Utility Theory) "synthesizing" the charac-

teristics of a given safety-critical system of interest and A^h the vulnerability class. In such a framework, a single system may be considered to belong to different vulnerability classes with different degrees of membership at the same time [132][133].

Further, the focus of this dissertation is concentrated on the improvement of given safety-critical systems. An inverse classification problem is solved to optimally choose a set of protective actions for each considered system in order to improve its performance (i.e., change its vulnerability class to a "desired" improved one). The optimization process relies on the choice of an objective function that is arbitrary and could depend on the given problem: e.g., imposing a set of "target" categories for each considered alternative, minimizing the cost of the chosen set of protective actions [86], ameliorating the "worst" alternative to a better category considering many possible compatible models etc. These are all interesting single-objective optimization problems to be explored. However, the need to face real applications renders the construction of a single objective function difficult and sometimes not feasible: the introduction of a multi-objective optimization framework would allow to manage more pieces of information and objectives at the same time. In our case, a model that considers simultaneously two or more objectives could produce solutions showing different trade-offs between, e.g., the cost of the actions and vulnerability of the systems [26]. In other words, the notion of Pareto optimality should be introduced [109][36][119]. Also, approaches such as the *scalarization* method [70], *ϵ-constraints* method [27], *Goal Programming*[29][28], *Multi-level Programming*[15][64][37][121] etc. could be considered.

Finally, it is worth noting that in the inverse classification problem (Chapter 4), for a given system it may be impossible to find out a set of protective actions that bring it to a "better" class. This may be due to its original characteristics (e.g., the physical distance of a NPP from a city can not be changed) and also to the set of available protective actions. Decision making usually focuses on the choice of a preferred solution among a set of *alternatives*. Typically, the decision maker concentrates first on the alternatives and only afterwards he/she addresses the objectives or criteria to evaluate such alternatives. This standard problem-solving approach is refered as

alternative-focused thinking. However, focusing on alternatives is a limited way to address decision-making situations. It is a reactive, not proactive attitude. It would make sense to have more control over the decision situations we face. This standard mode of thinking is backwards, because it puts the focus on identifying alternatives before articulating values. Instead, values are fundamentally important in any decision situation. Alternatives are relevant only because they are meant to achieve our expected values. Thus, our thinking should focus first on values and later on the alternatives that might achieve them [67][66]. This manner of thinking is refered to as *value-focused thinking* [65]. Based on the study and results presented in this dissertation, we have built a base to analyze the possible ways of ameliorating the systems in an alternative-focused way. The results have shown the limitations of the available and identified actions. With a value-focused thinking, instead, identifying or creating new decision alternatives (i.e., "new" protective actions in our case) to meet the goals and aspirations revealed through the MCDA process would be feasible. Also, it would be like a process of *identification of "better decision" situations*. These "better decision" situations, which we "create" for ourselves, should be thought of as *decision opportunities*, rather than as decision problems.

Bibliography

[1] C. C. Aggarwal, C. Chen, and J. W. Han. On the inverse classification problem and its applications. *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference. IEEE*, pages 111–113, 2006.

[2] C. C. Aggarwal, C. Chen, and J. W. Han. A framework for inverse classification. *Computer Science*, 2009.

[3] C. C. Aggarwal, C. Chen, and J. W. Han. The inverse classification problem. *JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, 25:458–468, 2010.

[4] K. Alsabti, S. Rank, and V. Singh. Clouds: A decision tree classifier for large datasets. In *KDD Conference*, page 2, 1998.

[5] G. E. Apostolakis, R. Piccinelli, and D.M. Lemon. A screening methodology for the identification and ranking of infrastructure vulnerabilities due to terrorism. *Risk Analysis*, 25:361–376, 2005.

[6] N. P. Archer and S. Wang. Application of the back propagation neural networks algorithm with monotonicity constraints for two-group classification problems. *Decision Sciences*, 24:60–75, 1993.

[7] T. Aven. *Misconceptions of Risk*. Wiley, Chichester, 2010.

[8] T. Aven and R. Flage. Use of decision criteria based on expected values to support decision-making in a production assurance and safety setting. *Reliability Engineering & System Safety*, 94:1491–1498, 2009.

[9] T. Aven and B. Heide. Reliability and validity of risk analysis. *Reliability Engineering & System Safety*, 94:1862–1868, 2009.

[10] P. Baraldi, R. Razavi-Far, and E. Zio. A method for estimating the confidence in the identification of nuclear transients by a bagged ensemble of fcm classifiers. *NPIC & HMIT*, pages 283–293, 2010.

[11] P. Baraldi, R. Razavi-Fra, and E. Zio. Bagged ensemble of fuzzy c means classifiers for nuclear transient identification. *Annals of Nuclear Energy*, 38:1161–1171, 2011.

[12] A. Bastian. Identifying fuzzy models utilizing genetic programming. *Fuzzy Sets and Systems*, 113:333–350, 2000.

[13] V. Belton and T. J. Stewart. *Multiple Criteria Decision Analysis: An Integrated Approach*. Kluwer Academic Publishers, Dordrecht, 2002.

[14] N. Ben-Asher and C. Gonzalez. Effects of cyber security knowledge on attack detection. *Computers in Human Behavior*, 48:51–61, 2015.

[15] W. Bialas and M. Karwan. Two-level linear programming. *Management Science*, 30:1004–1020, 1984.

[16] C. E. Bodley and P. E. Utgoff. Multivariate decision trees. *Machine Learning*, 20:63–94, 1995.

[17] G. Bous, P. Fortemps, F. Glineur, and M. Pirlot. ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements. *European Journal of Operational Research*, 206(2):435–444, 2010.

[18] D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, and P. Vincke. *Evaluation and Decision Models: Stepping Stones for the Analyst*. Springer Verlag, Berlin, 2006.

[19] D. Bouyssou, P. Perny, M. Pirlot, A. Tsoukià, and P. Vincke. A manifesto for the new mcda era. *Journal of Multi-Criteria Decision Analysis*, 2:125–127, 1993.

[20] J. P. Brans, B. Mareschal, and P. Vincke. How to select and how to rank projects: The promethee method. *European Journal of Operational Research*, 24:228–238, 1986.

[21] J. P. Brans and P. Vincke. A preference ranking organization method: the promethee method for multiple criteria decision-making. *Management Science*, 31:647–656, 1985.

[22] L. Breiman, J. Friedman, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[23] L. Breslow. Simplifying decision trees. *Knowledge Engineering Review*, 12:1–40, 1997.

[24] F. Cadini, E. Zio, V. Kopustinskas, and R. Urbonas. An empirical model based bootstrapped neural networks for computing the maximum fuel cladding temperature in a rbmk-1500 nuclear reactor accident. *Nuclear Engineering and Design*, 238:2165–2172, 2008.

[25] O. Cailloux and V. Mousseau. *Parameterize a territorial risk evaluation scale using multiple experts knowledge through risk assessment examples*, pages 2331–2339. Advances in Safety, Reliability and Risk Management. Taylor and Francis Group, London, 2011.

[26] M. Caramia and P. Dell'Olmo. *Multi-objective Optimization*, chapter 2. Multi-objective Management in Freight Logistics. Springer-Verlag, London, first edition, 2008.

[27] V. Chankong and Y. Y. Haimes. Multiobjective decision making. In *Theory and Methodology*, number 8 in North-Holland Series in System Science and Engineering. New York- Amsterdam: North- Holland Publishing Co, 1983.

[28] A. Charnes and W. W. Cooper. *Management Models and Industrial Applications of Linear Programming*. John Wile & Sons, 1961.

[29] A. Charnes, W. W. Cooper, and R. O. Ferguson. Optimalestimation of executive compensation by linear programming. *Management Science*, 2:138–151, 1955.

[30] R. Choo. The cyber threat landscape: Challenges and future research directions. *Computer & Security*, 30:719–731, 2011.

[31] P. J. Courtois. On time and space decomposition of complex structures. *Communications of the Acm*, 28:590–603, 1985.

[32] J. Depoy, Sandia Nat. Labs., N. M. Albuquerque, J. Phelan, P. Sholander, and B. Smith. Risk assessment for physical and cyber attacks on critical infrastructures. *Military Communications Conference*, 3:1961–1969, 2005.

[33] M. Doumpos and C. Zopounidis. *Multicriteria Decision Aid Classification Methods*. Kluwer Academic Publishers, Netherlands, 2002.

[34] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, New York, second edition, 2001.

[35] B. Efron and R.J. Thibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.

[36] M. Ehrgott. A discussion of scalarization techniques for multiple objective integer programming. *Annals of Operational Research*, 147:343–360, 2006.

[37] E. Erkut and F. Gzara. Solving the hazmat transport network design problem. *Computers and Operations Research*, 35:2234–2247, 2008.

[38] P. C. Fishburn. Independence in utility theory with whole product sets. *Operations Research*, 13:28–43, 1965.

[39] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[40] J. H. Friedman. A recursive partitioning decision rule for non-parametric classifiers. *IEEE Transactions on Computers*, pages 404–408, 1977.

[41] D. Furniss, J. Back, and A. Blandfod. A resilience markers framework for small teams. *Reliability Engineering & System Safety*, 96:2–10, 2011.

[42] J. Gehrke, V. Ganti, R. Ramakrishnan, and W. Y. Loh. Boat: Optimistic decision tree construction. *ACM SIGMOD Conference Proceedings*, 28:169–180, 1999.

[43] S. Gelfand, C. Ravishankar, and E. Delp. An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13/2:163–174, 1991.

[44] R. Gutierrez-Osuna. Pattern analysis for machine olfaction: A review. *IEEE SENSORS JOURNAL*, 10:189–202, 2002.

[45] Y. Y. Haimes. On the definition of resilience in systems. *Risk Analysis*, 29:198–501, 1974.

[46] Y. Y. Haimes. Modeling complex systems of systems with phantom system models. *Systems Engineering*, 15:333–346, 2012.

[47] M. Hofmann, G. H. Kjølle, and O. Gjerde. Development of indicators to monitor vulnerabilities in power systems. *2012 International Conference on Probabilistic Safety Assessment and Management (PSAM 11) & European Safety and RELiability Conference(ESREL 2012)*, 2012.

[48] J. P. Huang and K. L. Poh. Decision analysis in energy and environmental modeling. *Energy*, 20:843–855, 1995.

[49] P. Huard. *Resolution of mathematical programming with nonlinear constraints by the method of centers*, pages 209–219. Nonlinear Programming. Wiley, New York, 1967.

[50] M. S. Hung and J. W. Denton. Training neural networks with the grg2 nonlinear optimizer. *European Journal of Operational Research*, 69:83–91, 1993.

[51] A. V. Huzurbazar. Flowgraph models: a bayesian case study in construction engineering. *Journal of Statistical Planning and Inference*, 129:181–193, 2005.

[52] J. P. Ignizio. *Goal Programming and Extensions*. Lexington, 1972.

[53] J. P. Ignizio. Generalized goal programing: an overview. *Computers and Operations Research*, 10:277–289, 1985.

[54] M. Inuiguchi, T. Tanino, and M. Sakawa. Membership function elicitation in possibilistic programming problems. *Fuzzy Sets and Systems*, 111:29–45, 2000.

[55] B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. *Global sensitivity analysis*, 2014.

[56] H. Ishibuchi, K. Nozaki, and H. Tanaka. Distributed representation of fuzzy rules and its application to pattern classification. *Fuzzy Sets and Systems*, 52:21–32, 1992.

[57] H. Ishibuchi, K. Nozaki, and H. Tanaka. Efficient fuzzy partition of pattern space for classification problems. *Fuzzy Sets and Systems*, 59:295–304, 1993.

[58] E. Jacquet-Lagrèze. *Interactive assessment of preferences using holistic judgements: the PREFCALC system*, pages 335–350. Readings in Multiple Criteria Decision Aid. Springer Verlag, Berlin, 1990.

[59] E. Jacquet-Lagrèze and Y. Siskos. Preference disaggregation: 20 years of mcda experience. *European Journal of Operational Research*, 130:233–245, 2001.

[60] M. James. *Classification Algorithms*. Wiley, 1985.

[61] J. Johansson and H. Hassel. An approach for modelling interdependent infrastructures in the context of vulnerability analysis. *Reliability Engineering & System Safety*, 95:1335–1344, 2010.

[62] Quinlan J.R. chapter 4.5. Programs for Machine Learning. Morgan Kaufmann Publishers, California, 1993.

[63] K. Julisch. Understanding and overcoming cyber security anti-patterns. *Computers Networks*, 57:2206–2211, 2013.

[64] B. Y. Kara and V. Verter. Designing a road network for hazardous materials transportation. *Transportation Science*, 38:188–196, 2004.

[65] R. L. Keeney. Value-focused thinking: A path to creative decision making. *Journal of Multi-Criteria Decision Analysis*, 2:59, 1993.

[66] R. L. Keeney. *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press, 1996.

[67] R. L. Keeney. Value-focused thinking: Identifying decision opportunities and creating alternatives. *European Joumal of Operational Research*, 92:537–549, 1996.

[68] R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York, 1976.

[69] B. Kesler. The vulnerability of nuclear facilities to cyber attack. *Strategic Insights*, 10:15–25, 2011.

[70] I. Y. Kim and O. L. de Weck. Adaptive weighted sum method for multiobjective optimization: a new method for pareto front generation. *Structural and Multidisciplinary Optimization*, 31:105–116, 2006.

[71] T. Kim, S. Kwak, and S. Yoo. Applying multi-attribute utility theory to decision making in environmental planning: A case study of the electric utility in korea. *Journal of Environmental Planning and Management*, 41:597–609, 1998.

[72] Y. Kodratoff and R. S. Michalski. *Machine Learning: An Artificial Intelligence Approach*, volume 3. Morgan Kaufmann Publishers, California, 1990.

[73] A. M. Konce. Bulk power risk analysis: Ranking infrastructure elements according to their risk significance. *Electrical Power & Energy Systems*, 30:50–52, 2008.

[74] P. Korhonen and J. Wallenius. Behavioural issues in mcdm: neglected research questions. *Journal of Multi-Criteria Decision Analysis*, 5:178–182, 1996.

[75] B. Kosko. *Neural Networks and Fuzzy Systems*. Prentice-Hall, Englewood Cliffs, 1992.

[76] W. Kröger and E. Zio. *Vulnerable Systems*. Springer, UK, 2011.

[77] J. C. Laprie, K. Kanoun, and M. Kaaniche. Modelling interdependencies between the electricity and information infrastructures. *Computer Safety, Reliability, and Security*, pages 54–67, 2007.

[78] J. E. Larsson. *Knowledge-based methods for control systems*. PhD dissertation, Lund Institute of Technology, Department of Automatic Control, 1992.

[79] E. T. Lee and L. A. Zadeh. Note on fuzzy languages. *Information Sciences*, 1:421–434, 1969.

[80] S. M. Lee. *Goal Programming for Decision Analysis*. Auerbach, 1972.

[81] S. M. Lee and D. L. Olson. *Goal programming*, chapter 8. Multicriteria Decision Making. Springer US, 1999.

[82] A. Leroy, V Mousseau, and M. Pirlot. Learning the parameters of a multiple criteria sorting method based on a majority rule. *The Second International Conference on Algorithmic Decision Theory*, pages 219–233, 2001.

[83] A. G. Li, X. Zhou, and J. L. Zhang. Performance analysis of quantitative attributes inverse classification problem. *JOURNAL OF COMPUTERS*, 7, 2012.

[84] M. Lind. An introduction to multilevel flow modeling. *Nuclear safety and simulation*, 2:22–32, 2011.

[85] M. Lind. Reasoning about causes and consequences in multilevel flow models. *Advances in Safety, Reliability and Risk Management*, pages 2359–2367, 2011.

[86] M. V. Mannino and M. Koushik. The cost-minimizing inverse classification problem: a genetic algorithm approach. *Decision Support Systems*, 29:283–300, 2000.

[87] D. McFadden. *Conditional logit analysis in qualitative choice behavior*. Frontiers in Econometrics. Academic Press, New York, 1974.

[88] D. McFadden. *Structural discrete probability models derived from the theories of choice*. Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, 1980.

[89] M. F. Milazzo and T. Aven. An extended risk assessment approach for chemical plants applied to a study related to pipe ruptures. *Reliability Engineering & System Safety*, 94:183–192, 2012.

[90] M. Morgan, H. K. Florig, M. L. Dekay, and P. Fischbeck. Categorizing risks for risk ranking. *Risk Analysis*, 20:49–58, 2000.

[91] V. Mousseau, C. L. Dias, and J. Figueira. Dealing with inconsistent judgments in multiple criteria sorting models. *4OR: A Quarterly Journal of Operations Research*, 4:145–158, 2005.

[92] V. Mousseau and R. Slowinski. Inferring an electre tri model from assignment examples. *Journal of Global Optimization*, 12:157–174, 1998.

[93] National Water Resources Association NWRA. Risk assessment methods for water infrastructure systems. 2002.

[94] M. Ouyang. Review on modeling and simulation of interdependent critical infrastructure systems. *Reliability Engineering & System Safety*, 121:43–60, 2014.

[95] S. A. Patterson and G. E. Apostolakis. Identification of critical locations across multiple infrastructures for terrorist actions. *Reliability Engineering & System Safety*, 92:1183–1203, 2007.

[96] E. Patuwo, M. Y. Hu, and M. S. Hung. Two-group classification using neural networks. *Decision Sciences*, 24:825–845, 1993.

[97] P. C. Pendharkar. A potential use of data envelopment analysis for the inverse classification problem. *Omega*, pages 243–248, 2002.

[98] J. Piwowar, E. Chatelet, and P. Laclemence. An efficient process to reduce infrastructure vulnerabilities facing malevolence. *Reliability Engineering & System Safety*, 94:1869–1877, 2009.

[99] J. Pollet and J. Cummins. All-hazard approach for assessing readiness of critical infrastructure. *IEEE Conference on Technologies for Homeland Security, 2009*, pages 366–372, 2009.

[100] J. W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32:22–34, 1965.

[101] C. Raiu. Cyber-threat evolution: the past year. *Computer Fraud & Security*, 2012:5–8, 2012.

[102] C. Rocco and E. Zio. Bootstrap-based techiniques for computing confidence intervals in monte carlo system reliability evaluation. *Reliability and Maintainability Symposium*, pages 303–307, 1998.

[103] C. Romero. A survey of generalized goal programming. *European Journal of Operational Research*, 25:183–191, 1986.

[104] B. Roy. Classement et choix en présence de points de vue multiples: La méthode electre. *R.I.R.O*, 8:57–75, 1968.

[105] B. Roy. The outranking approach and the foundations of electre methods. *Theory and Decision*, 31:49–73, 1991.

[106] B. Roy. *Multicriteria Methodology for Decision Aiding*. Kluwer Academic Publishers, Dordrecht, 1996.

[107] B. Roy and D. Bouyssou. *Aide multicritère d'Aide à la Décision: Méthodes et Cas*. Economica, Paris, 1993.

[108] D. Ruan. *Logic-based hierarchies for modeling behavior of complex dynamic systems with applications*, volume 1 of *Fuzzy systems and soft computing in nuclear engineering*, chapter 17, pages 364–395. Springer-Verlag, 2000.

[109] S. Ruzika and M. Wiecek. Approximation methods in multiobjective programming. *Journal of Optimization Theory and Applications*, 3:473–501, 2005.

[110] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Carboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis*. Wiley, Chichester, 2008.

[111] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity analysis in practice*. Wiley, Chichester, 2004.

[112] M. J. Schniederjans. *Goal Programming: Methodology and Applications*. Cluwer, 1995.

[113] J. Shin, H. Son, R. Khalil, and G. Heo. Development of a cyber security risk model using bayesian networks. *Reliability Engineering & System Safety*, 134:208–217, 2014.

[114] Y. Siskos, E. Grigoroudis, and N. Matsatsinis. *UTA methods*, pages 297–344. Multiple Criteria Decision Analysis: State of the Art Surveys. Springer Verlag, Berlin, 2005.

[115] Y. Siskos, A. Spyridakos, and D. Yannacopoulos. Using artificial intelligence and visual techniques into preference disaggregation analysis: the miidas system. *European Journal of Operational Research*, 113:281–299, 1999.

[116] C. Smith. Some examples of discrimination. *Annals of Eugenics*, 13:272–282, 1947.

[117] G. Sonnevend. *An 'analytical centre' for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, pages 866–876. Lecture Notes in Control and Information Sciences. Springer Verlag, Berlin, 1985.

[118] V. Subramanian, M. S. Hung, and M. Y. Hu. An experimental evaluation of neural networks for classification. *Computers and Operations Research*, 20:769–782, 1993.

[119] V. T'Kindt and J. C. Billaut. *Multicriteria scheduling,theory,modelsand algorithms*. Springer-Verlag, Berlin, Heidelberg, 2005.

[120] G. W. Torrance and Boyle M. H. Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research*, 30:1043–1069, 1982.

[121] L. N. Vicente and P. H. Calamai. Bilevel and multilevel programming: a bibliography review. *Journal of Global Optimization*, 5:291–306, 1994.

[122] P. Vincke. *Multicriteria Decision-Aid*. John Wiley & Sons, Chichester, 1992.

[123] J. von Nuemann and O. Morgenstern. *Outranking approach*, chapter 8. Theory of Games and Linear Programming. Wiley, New York, second edition.

[124] D. von Winterfeldt and W. Edwards. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, 1986.

[125] T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio. Assessing the performance of a classification-based vulnerability analysis model. *Risk Analysis*, doi:10.1111/risa. 12305, 2014.

[126] T. R. Wang, V. Mousseau, and E. Zio. A hierarchical decision making framework for vulnerability analysis. *Proceedings of European Safety and RELiability Conference (ESREL 2013)*, pages 1–8, 2013.

[127] W. L. Waugh. Terrorism and the all-hazard model. *Journal of Emergency Management*, 4:8–10, 2005.

[128] G. F. White. Natural hazards, local, national, global. *Geographical Review*, 66:247–249, 1974.

[129] R. Wilson and T. R. Martinez. Combining cross-validation and confidence to measure fitness. *Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), 1999*, page 163, 1999.

[130] W. Yu. Electre tri: Aspects methodologiques et manuel d'utilisation. *Document du Lamsade, No.74*, 1992.

[131] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[132] L. A. Zadeh. Fuzzy algorithms. *Information and Control*, 12:94–102, 1968.

[133] L. A. Zadeh. Probability measures of fuzzy events. *Journal of mathematical Analysis and Applications*, 23:421–427, 1968.

[134] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning-i. *Information Sciencess*, 8:199–249, 1975.

[135] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning-ii. *Information Sciencess*, 8:301–357, 1975.

[136] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning-iii. *Information Sciencess*, 9:43–80, 1975.

[137] E. Zio. A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. *IEEE Transactions on Nuclear Science*, 53:1460–1470, 2006.

[138] E. Zio. *An Introduction to the Basics of Reliability and Risk Analysis*. World Scientific Publishing Co, 2007.

[139] E. Zio. Vulnerability and risk analysis of critical infrastructures. *Second International Conference on Vulnerability and Risk Analysis and Management (ICVRAM2014) and Sixth International Symposium on Uncertainty, Modeling and Analysis (ISUMA2014), University of Liverpool, UK.*, pages 23–30, 2014.

[140] E. Zio, R. Piccinelli, and G. Sansavini. An all-hazard approach for the vulnerability analysis of critical infrastructures. *ESREL 2011, Sep 2011, Troyes, France.*, pages 2451–2458, 2012.

[141] E. Zio, R. Piccinelli, and G. Sansavini. A framework for ranking the attack susceptibility of components of critical infrastructures. *The Italian Association of Chemical Engineering*, 26:117–122, 2012.

PART II
Appended papers

Paper (i)

A Hierarchical Decision Making Framework for Vulnerability Analysis.

T. R. Wang, V. Mousseau, and E. Zio.

Proceedings of the European Safety and RELiability Conference (ESREL2013), Sep 2013, Amsterdam, Netherlands. pp.1-8.

A Hierarchical Decision Making Framework for Vulnerability Analysis

T. R. WANG

Chair on Systems Science and the Energetic challenge, European Foundation for New Energy-Electricité de France, Ecole Centrale Paris and Supelec, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex

V. MOUSSEAU

Laboratory of Industrial Engineering, Ecole Centrale Paris, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex

E. ZIO

Politecnico di Milano, Energy Department, Nuclear Section, c/o Cesnef, via Ponzio 33/A , 20133, Milan, Italy
Chair on Systems Science and the Energetic challenge, European Foundation for New Energy-Electricité de France, Ecole Centrale Paris and Supelec, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex

ABSTRACT: Embracing an all-hazard view to deal with random failures, natural disasters, accidents and malevolent intentional acts, a framework for the vulnerability analysis of safety-critical systems and infrastructures is set up. A hierarchical structure is used to organise the information on the hazards, which is then manipulated through a decision-making process for vulnerability evaluation. We present the framework and its hierarchical model by way of assessing the susceptibility of a safety-critical system to intentional hazards, considering criteria of diverse nature, such as physical characteristics, social criticality characteristics, exposition to

cascading failures, resilience. We use a ranking method to compare systems of different characteristics. The systematic process of analysis is presented with reference to the exemplary case of nuclear power plants.

1 INTRODUCTION

The vulnerability of safety-critical systems and infrastructures is of great concern, given the multiple and diverse hazards that they are exposed to and the potential large-scale consequences.

We conceptualise vulnerability as a global system property related to the system susceptibility to all hazards, intentional, random internal and natural, and to resilience. Notably, resilience should not be considered separately but for its effects on the susceptibility to the three different kinds of hazards.

The susceptibility associated with random internal hazards and natural hazards is classically treated within a probabilistic framework to handle both the aleatory uncertainty in the occurrence of the accident events and their consequences (Kröger & Zio 2011) and the epistemic uncertainty on the hypotheses and parameters of the models used. Intentional hazards relate to malevolent acts and lack of a well-established methodology for accounting for uncertainty due to behaviours of different rationality (Depoy & Phelan 2005).

In this paper, we illustrate a decision-making framework intended to guide analysts, managers and stakeholders in the systematic identification of sources of vulnerability. Guided by the framework, effective management can be performed in an all-hazard perspective addressing questions like: what is the level of vulnerability of a site comparing with others? Which one should be protected and ameliorated? How to proceed and how much will it cost?

The evaluation through the framework is shown by way of analysing the susceptibility to intentional hazards of a safety-critical system, namely a Nuclear Power Plant (NPP), considering the vulnerability sources and the related features, the system technical and physical features, and the dependencies and interdependencies on other systems. The paper is organised as follows. In Section 2, the framework is presented with the focus on intentional hazards and by way of the reference example of NPPs. In Section 3, the decision-making methodology for assessing susceptibility is explained. In Section 4, an application is shown to exemplify the process. Conclusions are drawn in Section 5.

Table 1: Criteria, subcriteria and preference directions

Criterion	Physical characteristics	Social criticality	Possibility of cascading failures
Subcriteria	Number of workers	Percentage of contribution to the welfare	Connection distance
	Nominal power production	Size of served cities	
	Number of production units		
Preference direction	Min	Min	Min

Criterion	Recovery means	Human preparedness	Level of protection
Subcriteria	Number of installed backup components	Training	Physical size of the system
	Duration of backup components	Safety management	Number of accesses
	Duration of repair and recovery actions		Entrance control
	External emergency measures		Surveillance
Preference direction	Max	Max	Max

2 FRAMEWORK OF ANALYSIS

Vulnerability is defined in different ways depending on the domains of application, e.g.: vulnerability is a measure of possible future harm due to exposure to a hazard (Kröger & Zio 2011); the identification of weaknesses in security, focusing on defined threats that could compromise a system ability to provide a service (NWRA 2002); the set of conditions and processes resulting from physical, social, economic, and environmental factors, which increase the susceptibility of a community to the impact of hazards (Hofmann, Kjølle, & Gjerde 2012).

With the focus on the susceptibility to intentional hazards, a four-layers hierarchical model is shown in Figure 1. The susceptibility to intentional hazards is characterised in terms of attractiveness and accessibility. These are hierarchically broken down into factors which influence them, including resilience seen as pre-attack protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness). The decomposition is made in 6 criteria which are further decomposed in a layer of basic subcriteria, for which data and information can be collected to make their evaluation. The criteria and subcriteria considered serve as examples and are not to be considered exhaustive.

In the following subSections, the criteria of the layers are defined and assigned preference directions for treatment in the decision-making process. The preference direction of a criterion indicates towards which state it is desirable to lead it to reduce susceptibility, i.e., it is assigned from the point of view of the defender of an attack who is concerned with protecting the sys-

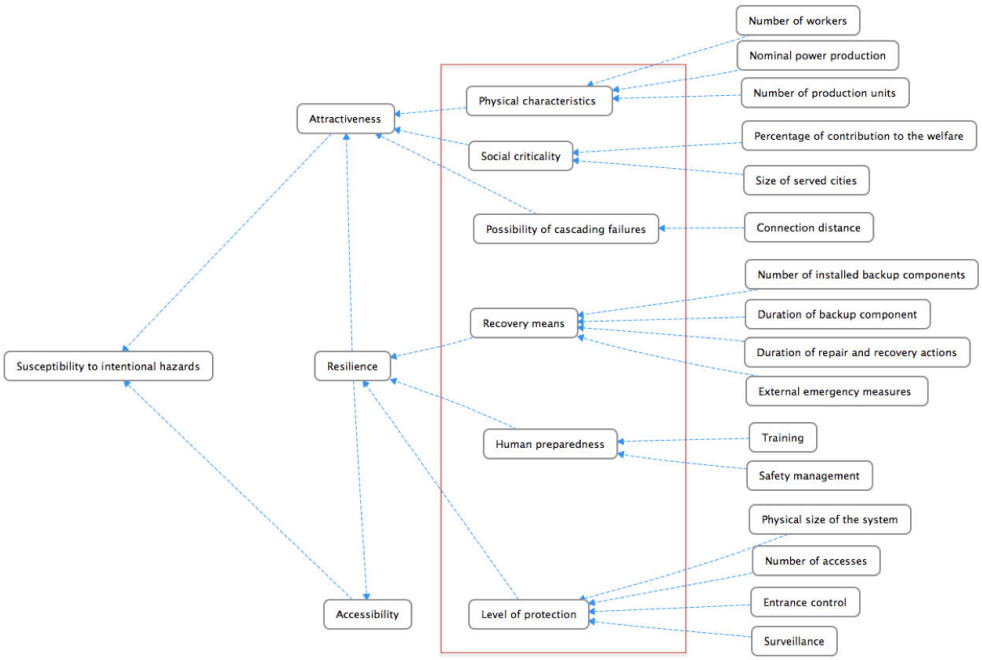

Figure 1: Hierarchical modelSusceptibility to intentional hazards

tem. Although only the 6 criteria in the third level of the hierarchy will be considered in the exemplary demonstration on the NPPs evaluation, examples of scales of evaluation also of the basic subcriteria of the last layer are proposed, in relation to the characteristics of NPPs for exemplification purposes.

2.1 *Attractiveness*

This second-layer criterion is intended to capture the interest that terrorists may have to attack the system. Such interest is considered to be driven mainly by the effects that the attack can cause, which include damages to the assets and environment, injured people, deaths. These depend on the physical characteristics of the system, its social criticality, the possibility of cascading effects and the system resilience. In a general sense, resilience represents the ability to avoid the occurrence of accidents despite the persistence of poor circumstances or to recover from some unexpected events (Furniss, Back, & Blandfod 2011). It is the ability of a system to anticipate, cope with/absorb, resist and recover from the impact of a hazard (technical) or a disaster (social). Resilience reflects a dynamic confluence of factors that promotes positive adaptation despite exposure to adverse life experiences. In our model, it is presented in terms of capacity of recovery, human preparedness and level of protection.

The preference direction characterising this factor is such that the more attractive the system is, the more it should be protected.

Table 2: Number of workers

Level	Number of workers
1	500
2	1000
3	1500
4	2000
5	2500

2.2 *Accessibility*

Accessibility is introduced as a criterion in the second layer of the hierarchy to describe the degree to which it is easy or difficult to arrive at a system in order to intentionally damage it. It is a function of resilience through the level of protection present to defend against malevolent attacks.

2.3 *Examples of subcriteria*

Each third-layer criterion is constituted by several subcriteria (Table 1). The value of the subcriteria can be crisp numbers or language terms according to the contents. Each of the subcriterion is analysed in giving an explanation of the contribution on the corresponding third-layer criterion.

2.3.1 *Number of workers*

This criterion can be seen to contribute to the attractiveness for an attack from various points of view, for example: 1) the more workers, the more work injuries and deaths from an attack; 2) the more workers, the easier for the attackers to sneak into the system; 3) the more workers, the higher the possibility that one of them can be turned into an attack. Limiting the number of workers can then contribute to the security of the plant and, thus, reduce its attractiveness for an attack. Table 2 reports some reference values, typical of NPPs.

2.3.2 *Entrance control*

This gives due count to the process and technology for entrance control. The more effective the control at the entrance is, the less easy it is to enter the system with bad intentions. Table 3 gives a 6 levels presentation.

Table 3: Entrance control

Level	Type of entrance control
1	Completely open, no control, no barriers
2	Unlocked, non-complex barriers
3	Complex barriers, security patrols
4	Secure area
5	Guarded, secure area, alarmed
6	Completely secure

3 ASSESSMENT METHODOLOGY

The hierarchical model just presented structures the susceptibility of a critical system to intentional attacks in terms of a number of criteria. The 16 basic, bottom-layer subcriteria are organised into 6 main ones: the physical characteristics, the social criticality, the possibility of cascading failures, the recovery means, the human preparedness and the level of protection. For the quantitative assessment, each of the 16 basic subcriteria needs to be assigned a value function in relation to the main criterion to which it contributes. The assignment can be done in relative terms, by comparing different systems with different characteristics. To exemplify how this is done, we consider NPPs as critical systems and perform a decision-making process for the evaluation of their characteristics with respect to susceptibility to intentional attacks. We first build a ranking of fictitious NPPs, through the authors' subjective preferential judgment of indirect data. This serves for constructing the basis for the relative evaluation of the characteristics of real NPPs.

To carry out the decision-making process for the evaluation, we resort to a multiple criteria decision aid (MCDA) technique named ACUTA (Analytic Centre UTilité Additive) based on the computation of the analytic centre of a polyhedron for the selection of additive value functions that are compatible with holistic assessments of the preferences in the criteria (Bous, Fortemps, Glineur, & Pirlot 2010). Being central by definition and uniquely defined, the analytic centre benefits from theoretical advantages over the notion of centrality used in other meta-UTA methods. A brief explanation of the method is given in the subSections that follow.

For the practical computations, we use an implementation of the method available in the Open Source software Diviz of the Decision Deck Project (http://www.decision deck.org/).

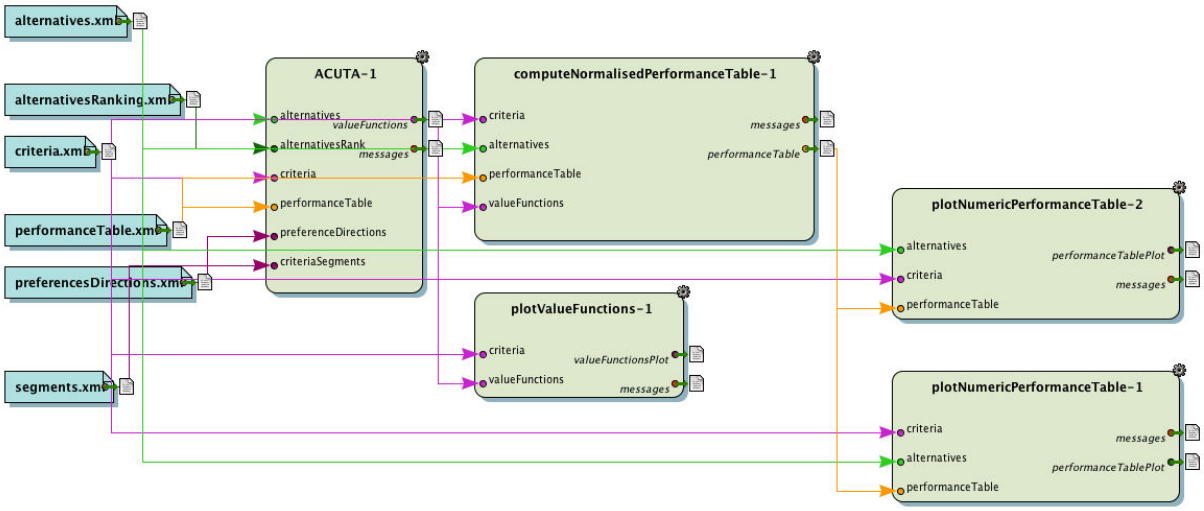

Figure 2: ACUTA analysis workflow for the illustrative example of Section 4

3.1 *Analytic Center*

The idea of the analytic centre of a polyhedron was first introduced by Huard (1967) and later reintroduced by Sonnevend (1985) in the context of convex optimization techniques. The theoretical framework around this concept lies at the heart of interior-point methods for solving linear programming optimisation problems. In ACUTA, it is suggested to compute a unique, well-defined and central solution for aggregation-disaggregation methods based on additive piecewise linear value function models (Bous, Fortemps, Glineur, & Pirlot 2010).

3.2 *ACUTA*

The UTA(UTilité Additive) method consists in building a piecewise linear additive decision model from a preference structure using linear programming. Let A be the set of possible alternatives and $A_L = \{a_j, j = 1, ..., k\}$ the learning set. In A_L, alternatives are ranked in order of decreasing preference by the DM (Decision Maker), i.e. $a_j \succsim a_{j+1}, j = 1, ..., k-1$, where $\succsim$ expresses that a_j is either preferred ($\succ$) or indifferent ($\sim$) to a_{j+1}. The values of the n criteria, denoted by $x_i (i = 1, ..., n)$, belong to the interval $[\overline{\chi_i}, \underline{\chi_i}]$ that, for each i, corresponds to the range between the worst ($\underline{\chi_i}$) and best $\overline{\chi_i}$ values found for attribute i among the alternatives in A. Our purpose is to establish marginal value functions $\nu_i(\chi_i)$ for each criterion in order to model the perceived value of each alternative. Since these values are piecewise linear functions, the range of values on each criterion is divided into subintervals using a predefined number of a_i points such that $\chi_i = \{\underline{\chi_i} = \chi_i^1, \chi_i^2, ..., \chi_i^{a_i} = \overline{\chi_i}\}$. The subdivision makes it possible to compute value functions by linear interpolation between the values $\nu_i(\chi_i^l)$ that have to be estimated and

hence appear as variables in the linear program. Using the degrees of freedom in the definition of a value function, we set $\nu_i(\underline{\chi_i}) = 0$ and

$$\sum_{i=1}^{n} \nu_i(\overline{\chi_i}) = 1 \tag{1}$$

This implies that $\nu_i(\overline{\chi_i})$ can be interpreted as the tradeoff associated to criterion i. Furthermore, all value functions should be monotonic, that is $\nu_i(\chi_i^{l+1}) - \nu_i(\chi_i^l) \geqslant \lambda(\forall\, i\ and\ l = 1, ..., a_i = 1)$, with $\lambda \geqslant 0$. According to the additive model, the global value $\nu(a_j)$ of an alternative a_j is given by the sum of its marginal values. In other terms, if the value of the j^{th} alternative on attribute i is denoted by a_{ij} , the global value of a_j is given by

$$\nu(a_j) = \sum_{i=1}^{n} \nu_i(a_{ij}) \tag{2}$$

This analytic expression of an alternative's global value allows for modelling the preferences of the DM, as expressed in the ranking of the learning set, using the following linear constraints, which we call preference constraints:

$$\nu(a_j) - \nu(a_{j+1}) \geqslant \delta \quad if\ a_j \succ a_{j+1}, \tag{3}$$

$$\nu(a_j) - \nu(a_{j+1}) = 0 \quad if\ a_j \sim a_{j+1}. \tag{4}$$

Here, λ is a positive number, called preference threshold, which is usually set to a small value. The assessment of the $\nu_i(\chi_i^l)$ variables should be done in such a way that the deviation from the preferences expressed by the DM in the subset A_L is minimal. The adaptation of the linear additive aggregation-disaggregation model to the analytic centre formulation is quite straightforward and gives rise to the ACUTA method; the introduction of slack variables into the objective function leads to the following nonlinear optimisation problem, which can be solved without further modifications:

$$max \sum_{j=1}^{k-1} ln(s_j) = \sum_{i=1}^{n} \sum_{l=1}^{a_i - 1} ln(s_{il}), \tag{5}$$

$$s.t.\ \ \nu(a_j) - \nu(a_{j+1}) = 0 \ \ if\ a_j \sim a_{j+1}, \tag{6}$$

$$(\nu(a_j) - \nu(a_j + 1)) - \delta = s_j \ \ if\ a_j \succ a_{j+1}, \tag{7}$$

$$s_{il} = (\nu(\chi_i^{l+1}) - \nu(\chi_i^l)) - \lambda, \tag{8}$$

$$\sum_{i=1}^{n} \nu_i(\overline{\chi_i}) = 1. \tag{9}$$

Since this approach maximises the sum of slacks, parameters δ and λ can be omitted, and this is considered an advantage. The essential advantage of this method, however, is the centrality and uniqueness of the solutions it produces.

3.3 *The Diviz tool*

Diviz is a software for designing, executing and sharing Multicriteria Decision Aid (MCDA) methods, algorithms and experiments. Based on basic algorithmic components, Diviz allows combining these criteria for creating complex MCDA workflows and methods.

Once the workflow is designed, it can be executed on various data sets written according to the XMCDA standard. This execution is performed on distant servers via web services (http://www.decision deck.org/diviz/).

Once the execution is completed, the outputs of the different elementary components are available and can be visualised in Diviz.

Figure 2 shows the workflow of the analysis of susceptibility to intentional hazards for the illustrative example on NPPs of Section 4. This workflow uses, among other components, the ACUTA component to determine value functions based on the ranking of the NPPs given by the authors. These value functions are then applied to the real plants values and the whole data is then analysed via some graphical representations.

4 ILLUSTRATIVE EXAMPLE

For illustration purposes, 9 fictitious plants are considered to obtain the value functions, which are in turn used to evaluate the susceptibility to intentional attacks of 9 real plants. In simple words, the former 9 fictitious plants are evaluated with respect to their susceptibility to intentional hazards, to build the base for comparison of the latter. Best (least vulnerable) and worst (most vulnerable) fictitious plants are defined as bounding references, by taking the best/worst conditions of all subcriteria considered. The details are presented in the following subSections.

4.1 *Case study preparation*

In the hierarchical model of susceptibility, we consider 16 basic subcriteria and 6 main criteria.

4.1.1 *Data preparation*

In order to apply the ACUTA method, a data preparation is necessary.

For the 9 fictitious plants (named F1 to F9), the data of the 16 subcriteria are assigned arbitrarily by the authors. The data of one basic subcriterion are assigned to the different fictitious sites in a way to ensure that all possible values of the subcriterion are included.

The worst (named fictitiousWorst) and best (named fictitousBest) fictitious plants are defined by taking the worst/best values of each basic subcriterion. These two fictitious plants bound, in worst and best, the situations that are expected from the other plants.

Then, the descriptive terms and values of the 16 subcriteria are scaled onto the categories.

To illustrate the procedure of comparison of the subcriteria, we refer to the level of the six aggregated main criteria introduced in Section 2 and listed in Table 1. Their preference directions are also presented. They convey the fact that it is preferable to limit the dimension of the plant, minimise social criticality, control the cascading failure, maximise the recovery means, give more training, be better prepared for emergency and take more protection measures.

To get the values of the six aggregated criteria, we apply a simple weighted sum to their constituents subcriteria. For this, the weights for each subcriterion are arbitrarily assigned by the authors. Then, the data of the 9 fictitious NPPs are normalised (that is, rescaled between 0 and 1).

Same steps are applied to the 9 real NPPs (named R1 to R9), whose data have been taken from publicly available documents.

The weights of each basic subcriterion in the group for the main criteria are the same as for the fictitious NPPs.

4.1.2 *Ranking of fictitious NPPs*

As presented in the previous Sections, the analysis using ACUTA method needs a ranking of the fictitious NPPs to begin with. It is usually given by the experts. In our case study, the utility functions are first given by the authors. As presented in Section 3.2, let N be the set of the 9 fictitious plants and $N_L = \{F_j, j = 1, ..., 9\}$ the learning set. The data of fictitiousWorst and fictitiousBest are used to be the limit interval for the given criterion, divided into 5 subintervals. The utility functions are given such that all the data of fictitiousWorst are set to 0 and the data sum of the fictitiousBest is set to 1. The value functions can be calculated and visualised in Figure 3.

Based on the utility functions of the main criteria and the data, we can obtain the marginal value of the corresponding criterion for each fictitious NPP.

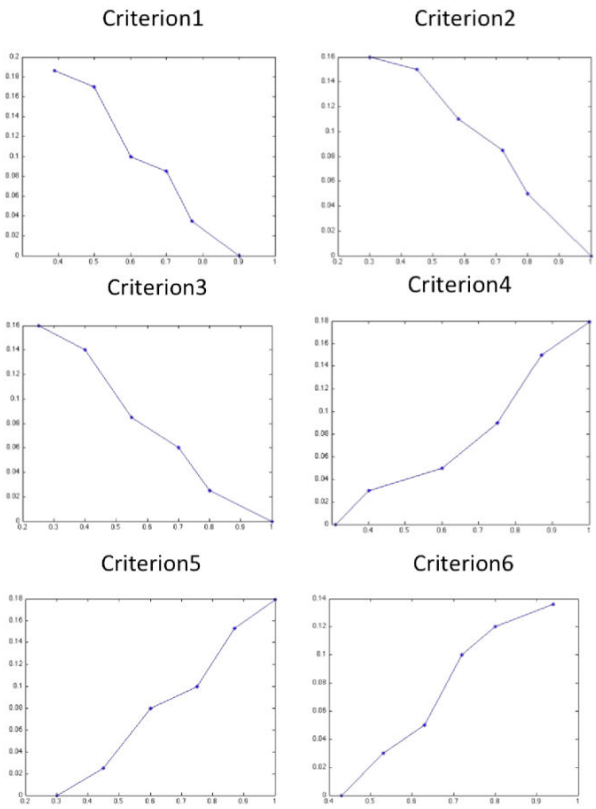

Figure 3: Utility functions given by authors

As a characteristic of the additive model, the global values which represent the susceptibility of the NPPs to the intentional hazards are given by the sum of its marginal values. These values are used to rank the NPPs. The ranking obtained is integrated into the decision-making process in the following subSection, to find out the value functions for the 6 criteria through the ACUTA method. The intentional hazards of real plants is then analysed and represented by using Diviz.

Table 4: Ranking of the fictitious NPPs based on the utility functions given by the authors

Rank	Name
1	fictitiousBest
2	F1
3	F2
4	F3
5	F4
6	F5
7	F6
8	F7
9	F8
10	F9
11	fictitiousWorst

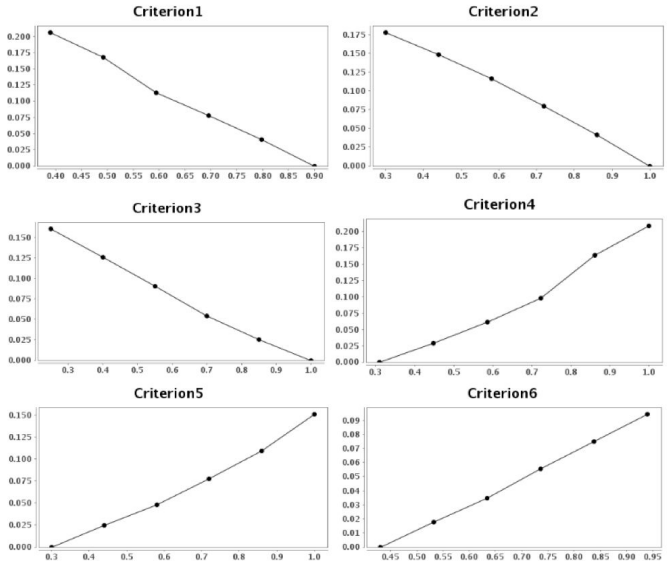

Figure 4: Representation of the Value Functions

4.2 *Results*

Applying the ACUTA method on the 9 fictitious plants in Diviz, we can calculate the value functions of the 6 criteria (Figure 4).

First of all, the criteria preference directions can be recognised easily from the trends of the curves. Also, for most part of each curve, it is natural that the vertical axis values are roughly proportional to the abscissa axis ones, because the vulnerability performance is roughly proportional to the value of the related parameters. More importantly, we can figure out the sensitive interval of each criterion. For example, for criterion 4, Recovery means, in the interval from 0.7 to 0.8 of the abscissa axis, there is an obvious change of gradient that is larger than before. This phenomenon also occurs in intervals of the other criteria and is due to the authors' preferences in the judgments. The more recovery means, the less susceptible the NPPs are to intentional hazards. Especially after a certain level (0.7 of the abscissa axis value), the extra-added measure can substantially increase the protection. This can be an indicator to know better the preference of the DMs during the ranking step and can also serve as a guidance during the amelioration of the plants.

In using the value functions, the former data of the 9 real NPPs can then be taken into account. We can compare the NPPs by single criterion. As shown in the 6 histograms (Figure 5), for one criterion, each column represents the corresponding performance of a given NPP. The length of each column is proportional to the marginal values. The longer the column, the better performance it has for the criterion. In the solid line frame there are the representative

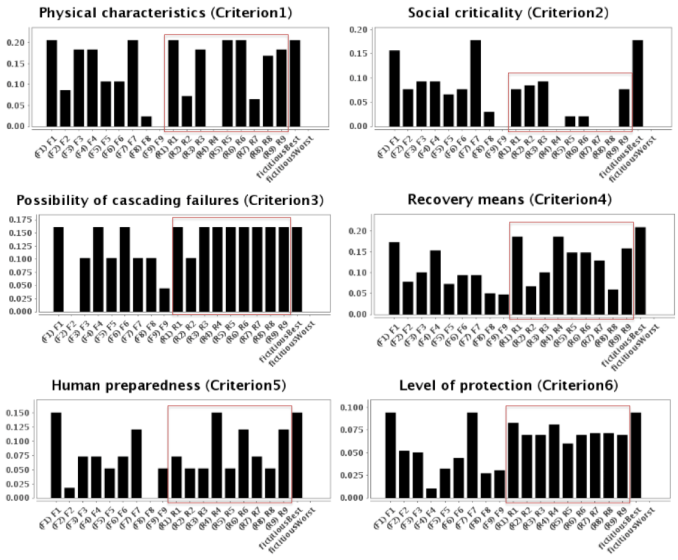

Figure 5: Histograms of subcriteria of the NPPs

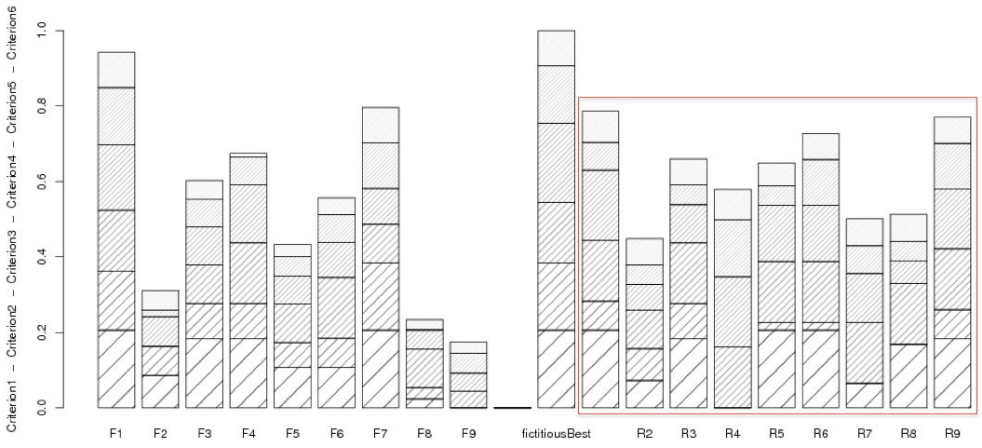

Figure 6: Histogram of susceptibility to intentional hazards of the NPPs

columns for each criterion of the real plants.

For the 6 criteria, the performances of most of the real NPPs are at least as good as the fictitious ones. Especially for possibility of cascading failure and level of protection, the performances of the real ones are nearly the best among all these 20 NPPs. But for the physical characteristics of the system, there are 3 plants that are worse than the others because of their higher production power and bigger size. For human preparedness, because of certain enhanced training and safety management systems, there are 3 plants that present a better result. For recovery means, the differences among the NPPs are not very big. And for the social criticality, they are more vulnerable than the fictitious ones.

As a characteristic of the additive model, the global values which represent the susceptibility of the NPPs to the intentional hazards are given by the sum of the marginal values. An overview

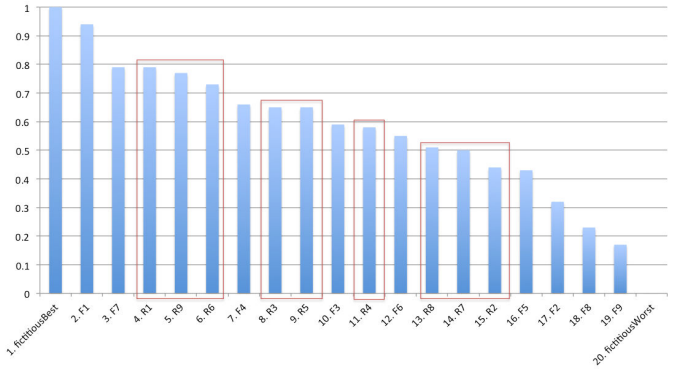

Figure 7: Ranking of the 20 NPPs

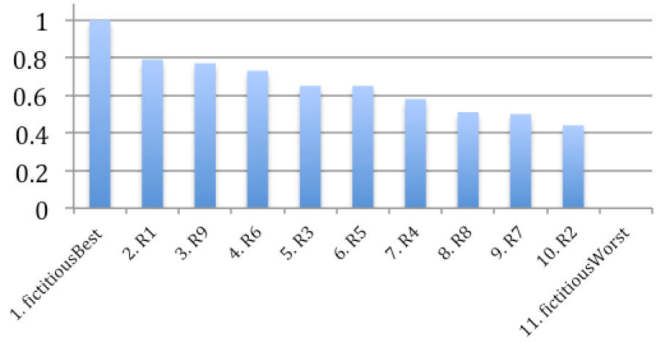

Figure 8: Ranking of the Real NPPs

of the 20 NPPs is presented graphically in Figure 6. Each column represents the susceptibility performance of one NPP to intentional hazards. Each column is constituted by 6 blocks with different textures that represent the 6 main criteria. As mentioned before, the height of each block of the representative column is proportional to the value of the corresponding criterion data. The smaller the height of the representative column of a plant is, the more susceptible it is in facing an intentional hazard.

Based on the performance values, we put the 20 NPPs in order as shown in Figure 7.

For the 9 real NPPs, according to the ranking, there are 5 that are among the first 10; all of them are among the first 15. Most of their performances are better than the fictitious ones. This is reasonable because for certain basic subcriteria, we have given some abnormally low values to the fictitious NPPs (e.g. for the basic subcriterion Type of entrance control, we have set certain fictitious plants to have unlocked barriers which is impossible for a real NPP). For the real NPPs, in view of production safety and international standards, certain criteria are already forced to be in limited intervals, that leads to improved situations than for the fictitious ones. We then concentrate only on the real NPPs, whose ranking result is given in Figure 8.

In order to find out the weaknesses of the real NPPs, we have done first the comparison

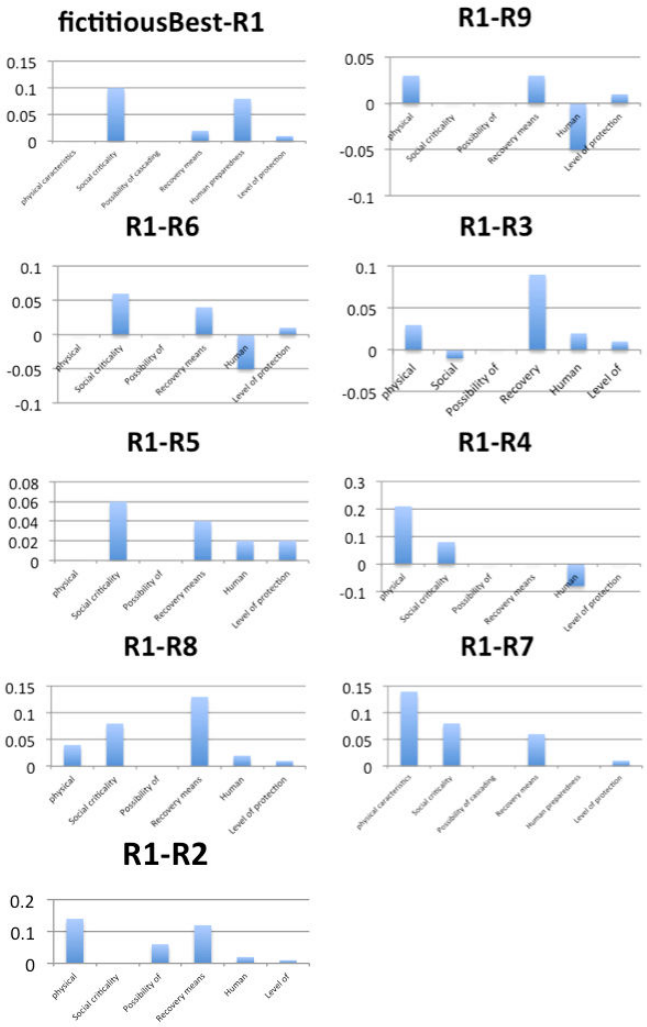

Figure 9: Performance comparison of the Real NPPs

between the fictitiousBest and R1 (which is the best among all the Real NPPs) and then between R1 and the rest of the Real NPPs, separately. The difference of the marginal value of each criterion is shown in Figure 9.

R1 is as good as the fictitiousBest for two criteria. In comparing with the rest of the Real NPPs, for R2, R5, R7 and R8, R1 is at least as good as them for each of the criteria. But for R3, R4, R6 and R9, R1 has an advantage only on the sum of the differences. There are criteria for which R1 is not as good as the others.

5 CONCLUSIONS

This paper proposes a decision making framework for analysing the vulnerability of critical infrastructures. A hierarchical model of susceptibility to intentional attacks has been taken as reference example. A case study of NPPs has been analysed by using the ACUTA method and the results calculated with the software Diviz.

The main contributions of this paper are the establishment of the hierarchical modelling framework for system vulnerability analysis and the decision making setting for its evaluation.

REFERENCES

Bous, G., P. Fortemps, F. Glineur, & M. Pirlot (2010). ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements. *European Journal of Operational Research 206*(2), 435–444.

Depoy, J. & J. Phelan (2005). Risk assessment for physical and cyber attacks on critical infrastructures. Volume 3.

Furniss, D., J. Back, & A. Blandfod (2011). A resilience markers framework for small teams. *Reliability Engineering & System Safety 96*.

Hofmann, M., G. Kjølle, & O. Gjerde (2012). Development of indicators to monitor vulnerabilities in power systems.

http://www.decision deck.org/.

http://www.decision deck.org/diviz/.

Huard, P. (1967). Resolution of mathematical programming with nonlinear constraints by the method of centers. *Nonlinear Programming*, 209–219.

Kröger, W. & E. Zio (2011). *Vulnerable Systems*. UK: Springer.

NWRA, N. W. R. A. (2002). Risk assessment methods for water infrastructure systems.

Sonnevend, G. (1985). An analytical centre for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming. *Lecture Notes in Control and Information Sciences*, 866–876.

Paper (ii)

Assessing the performance of a classification-based vulnerability analysis model

T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio.

Risk Analysis, Vol.35, No.9, 2015.

Risk Analysis, Vol. 35, No. 9, 2015 DOI: 0.1111/risa.12305

Assessing the Performance of a Classification-Based Vulnerability Analysis Model

Tairan Wang,[1⋆] Vincent Mousseau,[2] Nicola Pedroni,[1] Enrico Zio[1,3]

In this paper, a classification model based on the Majority Rule Sorting (MR-Sort) Method is employed to evaluate the vulnerability of safety-critical systems with respect to malevolent intentional acts. The model is built on the basis of a (limited-size) set of data representing (a priori-known) vulnerability classification examples.

The empirical construction of the classification model introduces a source of uncertainty into the vulnerability analysis process: a quantitative assessment of the performance of the classification model (in terms of accuracy and confidence in the assignments) is thus in order.

Three different approaches are here considered to this aim: (i) a model retrieval-based approach, (ii) the Bootstrap method and (iii) the Leave-one-out cross-validation technique. The analyses are presented with reference to an exemplificative case study involving the vulnerability assessment of nuclear power plants.

KEY WORDS: vulnerability analysis, classification model, confidence estimation, MR-Sort, nuclear power plants

[1]Chair on Systems Science and the Energy challenge, European Foundation for New Energy-Electricité de France, Ecole Centrale Paris and Supelec

[2]Laboratory of Industrial Engineering, Ecole Centrale Paris, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex, vincent.mousseau@ecp.fr

[3]Politecnico di Milano, Energy Department, Nuclear Section, c/o Cesnef, via Ponzio 33/A, 20133, Milan, Italy, enrico.zio@polimi.it

[⋆]Ecole Centrale Paris and Supelec, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex, tairan.wang@ecp.fr

1. INTRODUCTION

The vulnerability of safety-critical systems and infrastructures (e.g., nuclear power plants) is of great concern, given the multiple and diverse hazards that they are exposed to (e.g., intentional, random, natural etc.)[1] and the potential large-scale consequences. This has motivated an increased attention in analyses to guide designers, managers and stakeholders in (i) the systematic identification of the sources of vulnerability, (ii) its qualitative and quantitative assessment[2,3] and (iii) the selection of proper actions to reduce it. In this paper, we are concerned only with *intentional* hazards (i.e., those related to malevolent acts) and we mainly address issue (ii) above (i.e., the quantitative evaluation of vulnerability).

With respect to that, due to the specific features (low frequency but important effects) of intentional hazards (characterised by significant *uncertainties* due to behaviours of different rationality) the analysis is difficult to perform by traditional risk assessment methods[1,4,5]. For this reason, in the present work we propose to tackle the issue of evaluating vulnerability to malevolent intentional acts by an empirical classification modelling framework. In particular, we adopt a classification model based on the Majority Rule Sorting (MR-Sort) method[6] to assign an alternative of interest (i.e., a safety-critical system) to a given (vulnerability) class (or category). The MR-Sort classification model contains a group of (adjustable) parameters that have to be calibrated by means of a set of *empirical* classification examples (also called training set), i.e., a set of alternatives with the corresponding pre-assigned vulnerability classes.

Due to the finite (typically small) size of the set of training classification examples usually available in the analysis of real complex safety-critical systems, the performance of the classification model is impaired. In particular, (i) the classification *accuracy* (resp., error), i.e., the expected fraction of patterns correctly (resp., incorrectly) classified, is typically reduced (resp., increased); (ii) the classification process is characterised by significant uncertainty, which affects the *confidence* of the classification-based vulnerability model: in our work, we define the confidence in a classification assignment as in[10], i.e., as the probability that the class assigned by the model to a given (single) pattern is the correct one. Obviously, there is the possibility that a classification model assigns correctly a very large (expected) fraction of patterns (i.e., the model is very accurate), but at the same time *each* (correct) assignment is affected by significant uncertainty (i.e., it is characterised by low confidence). It is worth mentioning that besides the scarcity of training data, there are many additional sources of uncertainty in classification problems (e.g., the accuracy of the data, the suitability of the classification technique used, etc.): however, they are not considered in this work.

The performance of the classification model (i.e., the classification accuracy - resp., error - and the confidence in the classification) needs to be quantified: this is of paramount importance for taking robust decisions in the vulnerability analyses of safety-critical systems[7,8].

In this paper, three different approaches are used to assess the performance of a classification-based MR-Sort vulnerability model in the presence of small training data sets. The first is a model-retrieval based approach[6], which is used to assess the expected percentage error in assigning new alternatives. The second is based on *bootstrapping* the available training set in order to build an ensemble of vulnerability models[9]; the method can be used to assess both the accuracy and the confidence of the model: in particular, the confidence in the assignment of a given alternative is given in terms of the full (probability) distribution of the possible vulnerability classes for that alternative (built on the bootstrapped ensemble of vulnerability models)[10]. The third is based on the Leave-One-Out Cross-Validation

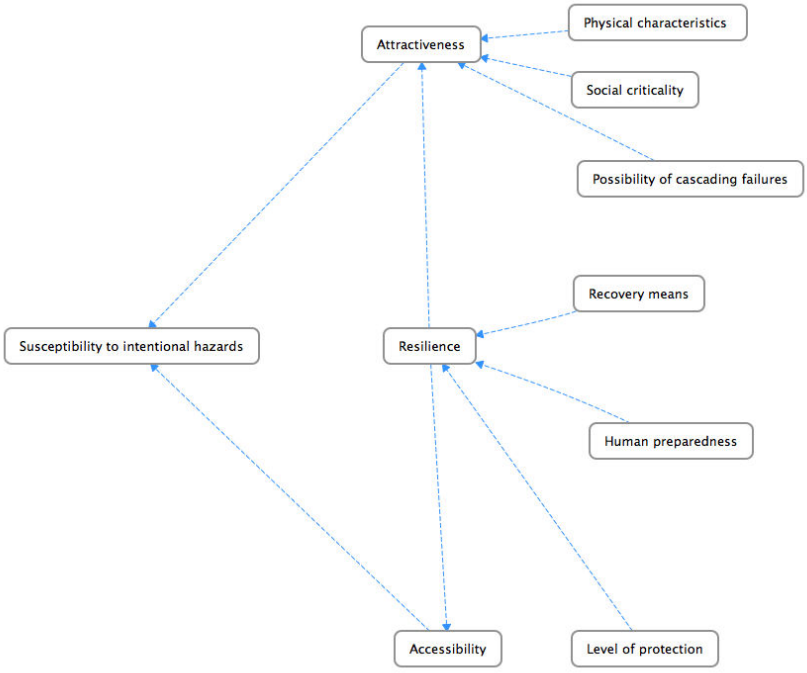

Fig. 1. Hierarchical model for susceptibility to intentional hazards.

technique, in which one element of the available data set is (left out and) used to test the accuracy of the classification model built on the remaining data: also this approach is employed to estimate the accuracy of the classification vulnerability model as the expected percentage error, i.e., the fraction of alternatives incorrectly assigned (computed as an average over the left-out data).

The contribution of this work is twofold:

- classification models have proved useful in a variety of fields including finance, marketing, environmental and energy management, human resources management, medicine, risk analysis, fault diagnosis etc.[(11)], but to the best of the authors? knowledge, this work is the first to propose a classification-based hierarchical framework for the analysis of the vulnerability to intentional hazards of safety-critical systems;
- the bootstrap method is originally applied to estimate the confidence in the assignments provided by the MR-Sort classification model, in terms of the probability that a given alternative is correctly classified.

The paper is organised as follows. The next Section presents the hierarchical framework for vulnerability analysis to intentional hazards. §3 shows the classification model applied within the proposed framework. §4 describes the learning process of a classification model by the disaggregation method. In §5, three approaches are proposed to analyse the performance of the classification model. Then, the proposed approaches are validated on the case study of a group of nuclear power plants in §6. Finally, §7 and §8 present the discussion and conclusions of this research.

2. GENERAL FRAMEWORK: VULNERABILITY TO INTENTIONAL HAZARDS

Vulnerability is defined in different ways depending on the domains of application, e.g., a measure of possible future harm due to exposure to a hazard [(1)], the identification of weaknesses in security, focusing on defined threats that could compromise a system ability to provide a service [(12)], the set of conditions and processes resulting from physical, social, economic, and environmental factors, which increase the susceptibility of a community to the impact of hazards [(13)].

With the focus on the susceptibility to intentional hazards, the three-layers hierarchical model developed in [(14)] is considered and shown in Figure 1. The susceptibility to intentional hazards is characterised in terms of attractiveness and accessibility. These are hierarchically broken down into factors which influence them, including resilience seen as pre-attack protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness). The decomposition is made in 6 criteria, which are further decomposed into a layer of basic sub-criteria, for which data and information can be collected. The details of the general framework of analysis are not given here for brevity; the interested reader is referred to [(14)] and to the Appendix A at the end of the paper.

For the purpose of the present paper, only six criteria are considered: physical characteristics, social criticality, possibility of cascading failures, recovery means, human preparedness and level of protection (Figure 1). These six criteria are used as the basis to assess the vulnerability of a given safety-critical system of interest (e.g., a nuclear power plant). Four levels (or categories) of vulnerability are considered: satisfactory, acceptable, problematic and serious. In this view, the issue of assessing vulnerability is here tackled within a classification framework: given the characterisation of a critical system in terms of the six criteria above, a proper vulnerability category (or class) has to be selected for that system. A description of the algorithm used to this purpose is given in the following Section.

It is worthy to mention that the cyber characteristics are not taken into account in this work; in future work they will be added for the criteria physical characteristics and protection.

3. CLASSIFICATION MODEL FOR VULNERABILITY ANALYSIS: THE MAJORITY RULE SORTING (MR-SORT) METHOD

The Majority Rule Sorting Model (MR-Sort) method is a simplified version of ELECTRE Tri, an outranking sorting procedure in which the assignment of an alternative to a given category is determined using a complex concordance non-discordance rule [(15) (16)]. We assume that the alternative to be classified (in this paper, a safety-critical system or infrastructure of interests, e.g., a nuclear power plant) can be described by an n-tuple of elements $x = \{x_1, x_2, ..., x_i, ..., x_n\}$, which represent the evaluation of the alternative with respect to a set of n criteria (by way of example, in the present paper the criteria used to evaluate the vulnerability of a safety critical system of interest may include its physical characteristics, social criticality, level of protection and so on: see §2). We denote the set of criteria by $N = \{1, 2, ..., i, ..., n\}$ and assume that the values x_i of criterion i range in the set X_i [(9)] (for example, in the present paper all the criteria range in [0, 1]). The MR-Sort procedure allows assigning any alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\} \in X = X_1 \times X_2 \times ... \times X_i \times ... \times X_n$ to a particular pre-defined category (in this paper, a class of vulnerability), in a given ordered set of categories, $\{A^h : h = 1, 2, ..., k\}$; as mentioned in §2, $k = 4$ categories are considered in this work: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, A^4 = serious.

To this aim, the model is further specialised in the following way:

- We assume that X_i is a subset of $\mathbb{R}$ for all $i \in \mathbb{N}$ and the sub-intervals $(X_i^1, X_i^2, ..., X_i^h, ..., X_i^k)$ of X_i are compatible with the order on the real numbers, i.e., for all $x_i^1 \in X_i^1, x_i^2 \in X_i^2, ..., x_i^h \in X_i^h, ..., x_i^k \in X_i^k$, we have $x_i^1 > x_i^2 > ... > x_i^h > ... > x_i^k$. We assume furthermore that each interval $x_i^h, h = 2, 3, ..., k$ has a smallest element b_i^h, which implies that $x_i^{h-1} \geq b_i^h > x_i^h$. The vector $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}$ (containing the lower bounds of the intervals X_i^h of criteria $i = 1, 2, ..., n$ in correspondence of category h) represents the lower limit profile of category A^h.
- There is a weight ω_i associated with each criterion $i = 1, 2, ..., n$, quantifying the relative importance of criterion i in the vulnerability assessment process; notice that the weights are normalised such that $\sum_{i=1}^{n} \omega_i = 1$.
  In this framework, a given alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\}$ is assigned to category $A^h, h = 1, 2, ..., k$, iff

$$\sum_{i \in \mathbb{N}: x_i \geq b_i^h} \omega_i \geq \lambda \; and \sum_{i \in \mathbb{N}: x_i \geq b_i^{h+1}} \omega_i < \lambda, \tag{1}$$

  where λ is a threshold ($0 \leq \lambda \leq 1$) chosen by the analyst. Rule (1) is interpreted as follows. An alternative x belongs to category A^h if: 1) its evaluations in correspondence of the n criteria (i.e., the values $\{x_1, x_2, ..., x_i, ..., x_n\}$) are at least as good as b_i^h(lower limit of category A^h with respect to criterion i), $i = 1, 2, ..., n$, on a subset of criteria that has sufficient importance (in other words, on a subset of criteria that has a weight larger than or equal to the threshold λ chosen by the analyst); and at the same time 2) the weight of the subset of criteria on which the evaluations $\{x_1, x_2, ..., x_i, ..., x_n\}$ are at least as good as b_i^{h+1}(lower limit of the successive category A^{h+1} with respect to criterion i), $i = 1, 2, ..., n$, is not sufficient to justify the assignment of x to the successive category A^{h+1}.
  Notice that alternative x is assigned to the best category A^1 if $\sum_{i \in \mathbb{N}: x_i \geq b_i^1} \omega_i \geq \lambda$ and it is assigned to the worst category A_k if $\sum_{i \in \mathbb{N}: x_i \geq b^{-k-1}} \omega_i < \lambda$. Finally, it is straightforward to notice that the parameters of such a model are the $k \cdot n$ lower limit profiles (n limits for each of the k categories), the n weights of the criteria $\omega_1, \omega_2, ..., \omega_i, ..., \omega_n$, and the threshold λ, for a total of *n(k+1)+1* parameters.

4. CONSTRUCTING THE MR-SORT CLASSIFICATION MODEL

In order to construct an MR-Sort classification model, we need to determine the set of *n(k+1)+1* parameters described in the previous §2, i.e., the weights $\omega = \{\omega_1, \omega_2, ..., \omega_n\}$, the lower profiles $b = \{b^1, b^2, ..., b^h, ..., b^k\}$,with $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}, h = 1, 2, ..., k$, and the threshold λ; in this paper, λ is considered a fixed, constant value chosen by the analyst (e.g., λ=0.9).

To this aim, the decision maker provides a training set of ?classification examples? $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{TR}\}$, i.e., a set of N_{TR} alternatives (in this case, nuclear power plants) $x_p = \{x_1^p, x_2^p, ..., x_i^p, ..., x_n^p\}, p = 1, 2, ..., N_{TR}$ together with the corresponding real pre-assigned categories (i.e., vulnerability classes) Γ_p^t (the superscript ?t? indicates that Γ_p^t represents the true, a priori-known vulnerability class of alternative x_p).

The calibration of the *n(k+1)* parameters is done through the learning process detailed in [6]. In extreme synthesis, the information contained in the training set D_{TR} is used to restrict the set of MR-Sort models compatible with such information, and to finally select one among them [6]. The a priori-known assignments generate constraints on the parameters of the MR-Sort model. In [6], such constraints have a linear formulation and are integrated into a Mixed

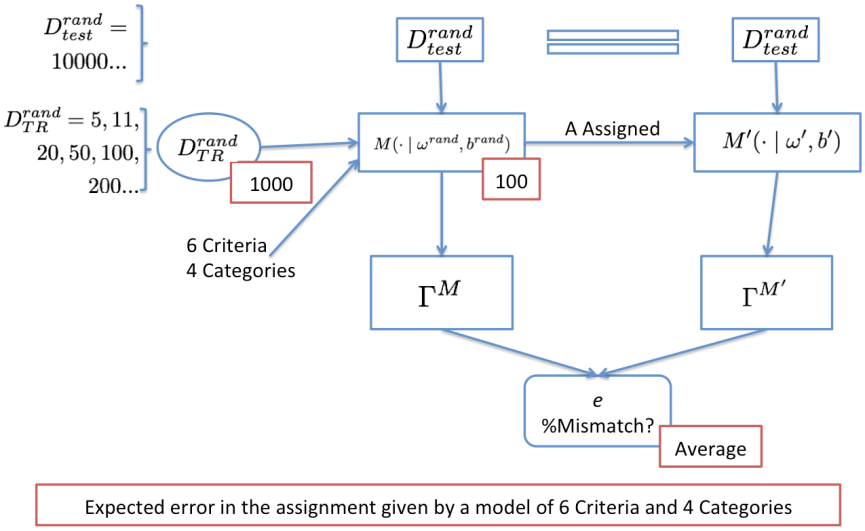

Fig. 2. The general structure of the model-retrieval approach.

Integer Program (MIP) that is designed to select one (optimal) set of such parameters ω^* and b^* (in other words, to select one classification model $M(\cdot|\omega^*, b^*)$) that is coherent with the data available and maximises a defined *objective function*. In [6], the optimal parameters ω^* and b^* are those that maximise the value of the minimal slack in the constraints generated by the given set of data D_{TR}. Once the (optimal) classification model $M(\cdot|\omega^*, b^*)$ is constructed, it can be used to assign a new alternative x(i.e., a new nuclear power plant) to one of the vulnerability classes $A^h, h = 1, 2, ..., k$: in other words, $M(x|\omega^*, b^*) = \Gamma^M_x$ where Γ^M_x is the class assigned by model $M(\cdot|\omega^*, b^*)$ to alternative x and assumes one value among $\{A^h : h = 1, 2, ..., k\}$. Further mathematical details about the training algorithm are not given here for brevity: the reader is referred to [6] and to the Appendix B at the end of the paper.

Obviously, the number N_{TR} of available classification examples is finite and quite small, in most of real applications involving the vulnerability analysis of safety-critical systems. As a consequence, the model $M(\cdot|\omega^*, b^*)$ is only a partial representation of reality and its assignments are affected by uncertainty: this uncertainty, which needs to be quantified to build confidence in the decision process, which follows the vulnerability assessment.

In the following Section, three different methods are presented to assess the performance of the MR-sort classification model.

5. METHODS FOR ASSESSING THE PERFORMANCE OF THE CLASSIFICATION-BASED VULNERABILITY ANALYSIS MODEL

5.1 Model Retrieval-Based Approach

The first method is based on the model-retrieval approach proposed in [6]. A fictitious set D^{rand}_{TR} of N_{TR} alternatives $\{x^{rand}_p : p = 1, 2, ..., N_{TR}\}$ is generated by random sampling within the ranges X_i of the criteria, $i = 1, 2, ..., n$. Notice that the size N_{TR} of the fictitious set D^{rand}_{TR} has to be the same as the real training set D_{TR} available, for the comparison to be fair. Also, a MR-Sort classification model $M(\cdot|\omega^{rand}, b^{rand})$ is constructed by randomly sampling possible values

of the internal parameters, $\{\omega_i : i = 1, 2, ..., n\}$ and $\{b_h : h = 1, 2, ..., k-1\}$. Then, we simulate the behaviour of a Decision Maker (DM) by letting the (random) model $M(\cdot|\omega^{rand}, b^{rand})$ assign the (randomly generated) alternatives $\{x_p^{rand} : p = 1, 2, ..., N_{TR}\}$. In other words, we construct a learning set D_{TR}^{rand} by assigning the (randomly generated) alternatives using the (randomly generated) MR-Sort model, i.e., $D_{TR}^{rand} = \{(x_p^{rand}, \Gamma_p^M) : p = 1, 2, ..., N_{TR}\}$, where Γ_p^M is the class assigned by model $M(\cdot|\omega^{rand}, b^{rand})$ to alternative x_p^{rand}, i.e., $\Gamma_p^M = M(x_p^{rand}|\omega^{rand}, b^{rand})$. Subsequently, a new MR-Sort model $M'(\cdot|\omega', b')$, compatible with the training set D_{TR}^{rand}, is inferred using the MIP formulation summarised in §3 and in the Appendix B. Although models $M(\cdot|\omega^{rand}, b^{rand})$ and $M'(\cdot|\omega', b')$ may be quite different, they coincide on the way they assign elements of D_{TR}^{rand}, by construction. In order to compare models M and M', we randomly generate a (typically large) set D_{test}^{rand} of *new* alternatives $D_{test}^{rand} = \{x_p^{test,rand} : p = 1, 2, ..., N_{Test}\}$ and we compute the percentage of ?assignment errors?, i.e., the proportion of these N_{Test} alternatives that models M and M' assign to different categories.

In order to account for the randomness in the generation of the training set D_{TR}^{rand} and of the model $M(\cdot|\omega^{rand}, b^{rand})$, and to provide robust estimates for the assignment errors ϵ, the procedure outlined above is repeated for a large number N_{sets} of random training sets $D_{TR}^{rand,j}, j = 1, 2, ..., N_{sets}$; in addition, for each set j the procedure is repeated for different random models $M(\cdot|\omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models}$. The sequence of assignment errors thereby generated, $e_{jl}, j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$, is then averaged to obtain a robust estimate for ϵ. The procedure is sketched in Figure 2.

Notice that this method does not make any use of the original training set D_{TR}(i.e., of the training set constituted by real-world classification examples). In this view, the model retrieval-based approach can be interpreted as a tool to obtain an absolute evaluation of the expected error that an 'average' MR-Sort classification model $M(\cdot|\omega, b)$ with k categories, n criteria and trained by means of an 'average' data set of given size N_{TR} makes in the task of classifying a new generic (unknown) alternative.

5.2 The Bootstrap Method

A way to assess *both* the accuracy (i.e., the expected fraction of alternatives correctly classified) *and* the confidence of the classification model (i.e., the probability that the category assigned to a given alternative is the correct one) is by resorting to the bootstrap method [17], which is used to create an ensemble of classification models constructed on different data sets bootstrapped from the original one [18]: the final class assignment provided by the ensemble is based on the combination of the individual output of classes provided by the ensemble of models [10].

The basic idea is to generate different training datasets by random sampling with replacement from the original one [17]: such different training sets are used to build different individual classification models of the ensemble. In this way, the individual classifiers of the ensemble possibly perform well in different regions of the training space and thus they are expected to make errors on alternatives with different characteristics; these errors are balanced out in the combination, so that the performance of the ensemble of bootstrapped classification models is in general superior than that of the single classifiers [18] [19]. This is a desirable property since it is a more realistic simulation of the real-life experiment from which our dataset was obtained. In this paper, the output classes of the single classifiers are combined by *majority voting*: the class chosen by most classifiers is the ensemble assignment. Finally, the accuracy of the model is

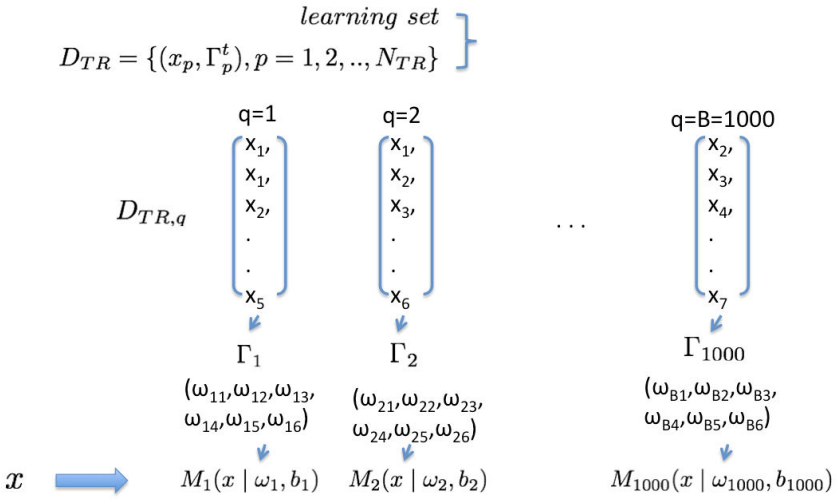

Fig. 3. The bootstrap algorithm.

given by the fraction of the patterns correctly classified. The bootstrap-based empirical distribution of the assignments given by the different classification models of the ensemble is then used to measure the confidence in the classification of a given alternative x that represent the probability that this alternative is correctly assigned [(10) (20)].

In more detail, the main steps of the bootstrap algorithm are as follows (Figure 3):

1. Build an ensemble of B (typically of the order of 500-1000) classification models $\{M_q(\cdot|(\omega_q, b_q) : q = 1, 2, ..., B)\}$ by random sampling with replacement from the original data set D_{TR} and use each of the bootstrapped models $M_q(\cdot|\omega_q, b_q)$ to assign a class $\Gamma_x^q, q = 1, 2, ..., B$, to a given alternative x of interest (notice that Γ_x^q takes a value in $A^h, h = 1, 2, ..., k$). By so doing, a bootstrap-based empirical probability distribution $P(A^h|x), h = 1, 2, ..., k$ for category A^h of alternative x is produced, which is the basis for assessing the confidence in the assignment of alternative x. In particular, repeat the following steps for $q = 1, 2, ..., B$:
    a. Generate a bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$, by performing random sampling with replacement from the original data set $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$ of N_{TR} input/output patterns. The data set $D_{TR,q}$ is thus constituted by the same number N_{TR} of input/output patterns drawn among those in D_{TR}, although due to the sampling with replacement some of the patterns in D_{TR} will appear more than once in $D_{TR,q}$, whereas some will not appear at all.
    b. Build a classification model $\{M_q(\cdot|\omega_q, b_q) : q = 1, 2, ..., B\}$, on the basis of the bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$.
    c. Use the classification model $M_q(\cdot|\omega_q, b_q)$ to provide a class $\Gamma_x^q, q = 1, 2, ..., B$ to a given alternative of interest, i.e., $\Gamma_x^q = M_q(x|\omega_q, b_q)$.
2. Combine the output classes $\Gamma^q, q = 1, 2, ..., B$ of the individual classifiers by majority voting: the class chosen by most classifiers is the ensemble assignment $\Gamma_x^{ens}, i.e.,\ \Gamma_x^{ens} = argmax_{A^h}[card_q\{\Gamma_x^q = A^h\}]$.
3. As an estimation of the confidence in the majority-voting assignment Γ_x^{ens}(step 2, above), we consider the bootstrap-based empirical probability distribution $P(A^h|x), h = 1, 2, ..., k$, i.e., the probability that category

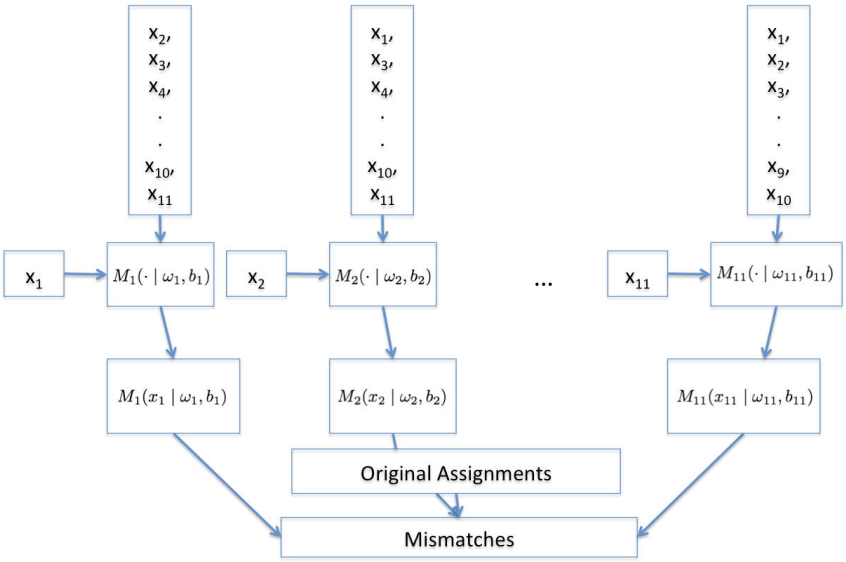

Fig. 4. Leave-one-out Cross-Validation Study procedure.

A^h is the correct category given that the (test) alternative is x [6]. The estimator of $P(A^h|x)$ here employed is: $P(A^h|x) = \frac{\sum_{q=1}^{B} I\{\Gamma_q = A^h\}}{B}$, where $I\{\Gamma^q = A^h\} = 1$, if $\Gamma_q = A^h$, and 0 otherwise.

4. Finally, the error of classification is presented by the fraction of the number of the alternatives being assigned by the classification model and the total number of the alternatives. The accuracy of the classification model is defined as the complement to 1 to the error.

5.3 The Leave-One-Out Cross-Validation (LOOCV) Technique

Leave-One-Out Cross-Validation (LOOCV) is a particular case of the cross-validation method. In cross-validation, the original training set D_{TR} is divided into N partitions, $A_1, A_2, ..., A_N$, and the elements in each of the partitions are classified by a model trained by means of the elements in the remaining partitions (Leave-p-out Cross-Validation) [20]. The cross-validation error is, then, the average of the N individual error estimates. When N is equal to the number of elements N_{TR} in D_{TR}, the result is Leave-One-Out Cross-Validation (LOOCV), in which each instance $x_p, p = 1, 2, ..., N_{TR}$ is classified by all the instances in D_{TR} except for itself [21]. For each instance $x_p, p = 1, 2, ..., N_{TR}$ in D_{TR}, the classification accuracy is *1* if the element is classified correctly and *0* if it is not. Thus, the average LOOCV error (resp. accuracy) over all the N_{TR} instances in D_{TR} is ϵ/N_{TR}(resp. $1 - \epsilon/N_{TR}$), where ϵ(resp. $N_{TR} - \epsilon$) is the number of elements incorrectly (resp. correctly) classified. Thus, the accuracy in the assignment is estimated as $1 - \epsilon/N_{TR}$.

With respect to the Leave-p-Out Cross-Validation, the Leave-One-Out Cross-Validation (LOOCV) produces a smaller bias of the true error rate estimator. However, the computational time increases significantly with the size of the data set available. This is the reason why the LOOCV is particularly useful in the case of small data sets. In addition, for *very sparse* datasets (e.g., of size lower than or equal to ten), we may be "*forced*" to use LOOCV in order to maximise the number of training examples employed and to generate training sets containing an amount of information that is

Table I . Training set with $N_{TR} = 11$ assigned alternatives

Alternatives, x_p	Vulnerability Class Γ_p^t
$x_1 = \{0.61, 0.6, 0.75, 0.86, 1, 0.94\}$	A^1
$x_2 = \{0.33, 0.27, 0, 0.575, 0.4, 0.72\}$	A^3
$x_3 = \{0.55, 0.33, 0.5, 0.725, 0.7, 0.71\}$	A^2
$x_4 = \{0.55, 0.33, 0.75, 0.8, 0.7, 0.49\}$	A^3
$x_5 = \{0.39, 0.23, 0.5, 0.6, 0.6, 0.62\}$	A^3
$x_6 = \{0.39, 0.27, 0.75, 0.725, 0.7, 0.68\}$	A^2
$x_7 = \{0.61, 0.7, 0.5, 0.725, 0.9, 0.94\}$	A^2
$x_8 = \{0.16, 0.1, 0.5, 0.475, 0.3, 0.59\}$	A^4
$x_9 = \{0.1, 0, 0.25, 0.5, 0.6, 0.61\}$	A^4
$x_{10} = \{0.1, 0, 0, 0.3, 0.3, 0.43\}$	A^4
$x_{11} = \{0.61, 0.7, 0.75, 1, 1, 0.94\}$	A^1

sufficient and reasonable for building an empirical model[(22)]. In Figure 4, the algorithm is sketched with reference to a training set D_{TR} containing $N_{TR} = 11$ data (like in the case study considered in the following §6).

6. APPLICATION

The methods presented in §5 are here applied on an exemplificative case study concerning the vulnerability analysis of Nuclear Power Plants (NPPs)[(?)]. We identify $n = 6$ main criteria $i = 1, 2, ..., n = 6$ by means of the hierarchical approach presented in[(?)], see Chapter 3: x_1 = physical characteristics, x_2 = social criticality, x_3 = possibility of cascading failures, x_4 = recovery means, x_5 = human preparedness and x_6 = level of protection. Then, $k = 4$ vulnerability categories $A^h, h = 1, 2, ..., k = 4$ are defined as: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic and A^4 = serious (Chapter 3). The training set D_{TR} is constituted by a group of $N_{TR} = 11$ NPPs x_p with the corresponding a priori-known categories Γ_p^t, i.e., $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR} = 11\}$. The training set is summarised in Table I.

In what follows, the three techniques of §5 are applied to assess the performance of the MR-Sort classification-based vulnerability analysis model built using the training set D_{TR} of Table I.

6.1 Application of the Model Retrieval-Based Approach

We generate $N_{sets} = 1000$ different training sets $D_{TR}^{rand,j}, j = 1, 2, ..., N_{sets}$, and for each set j, we randomly generate $N_{models} = 100$ models $M(\cdot|\omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models} = 100$. By so doing, the expected accuracy (1-ϵ) of the corresponding MR-Sort model is obtained as the average of $N_{sets} \cdot N_{models} = 1000 \cdot 100 = 100000$ values $(1 - \epsilon_{jl}), j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$(see §5.1). The size N_{test} of the random test set D_{TR}^{rand} is $N_{test} = 10000$. Finally, we perform the procedure of §5.1 for different sizes N_{TR} of the random training set D_{TR}^{rand}(even if the size of the real training set available is $N_{TR} = 11$, see Table I): in particular, we choose $N_{TR} = 5, 11, 20, 50, 100$ *and* 200. This

analysis serves the purpose of outlining the behaviour of the accuracy $(1-\epsilon)$ as a function of the amount of classification examples available.

The results are summarised in Figure 5 where the average percentage assignment error ϵ is shown as a function of the size N_{TR} of the learning set (from 5 to 200). As expected, the assignment error ϵ tends to decrease when the size of the learning set N_{TR} increases: the higher the cardinality of the learning set, the higher (resp. lower) the accuracy (resp. the expected error) in the corresponding assignments. Comparing these results with those obtained by Leroy et al[6] using MR-Sort models with $k = 2$ and 3 categories and n = 3-5 criteria, it can be seen that for a given size of the learning set, the error rate (resp. the accuracy) grows (resp. decrease) with the number of model parameters to be determined by the training algorithm = *n(k+1)+1*. It can be seen that for our model with n = 6 criteria and k = 4 categories, in order to guarantee an error rate inferior to 10% we would need training sets consisting of more than N_{TR} = 100 alternatives. Typically, for a learning set of $N_{TR} = 11$ alternatives (like that available in the present case study), the average assignment error ϵ is around 30%; correspondingly, the accuracy of the MR-Sort classification model trained with the data set D_{TR} of size $N_{TR} = 11$ available in the present case is around $(1 - \epsilon) = 70\%$: in other words, there is a probability of 70% that a new alternative (i.e., a new NPP) is assigned to the correct category of vulnerability.

In order to assess the randomness intrinsic in the procedure used to obtain the accuracy estimate above, we have also calculated the 95% confidence intervals for the average assignment error ϵ of the models trained with $N_{TR} = 11, 20$ *and* 100 alternatives in the training set. The 95% confidence interval for the error associated to the models trained with 11, 20 and 100 alternatives as learning set are [25.4%, 33%], [22.2%, 29.3%] and [10%, 15.5%], respectively. For illustration purposes, Figure 6 shows the distribution of the assignment mismatch built using the $N_{sets} \cdot N_{models} = 100000$ values $\epsilon_{jl}, j = 1, 2, ..., N_{sets} = 1000, l = 1, 2, ..., N_{models} = 100$, generated as described in §5.1 for the example of 11 alternatives.

6.2 Application of The Bootstrap Method

A number B (= 1000) of bootstrapped training sets $D_{TR,q}, q = 1, 2, ..., 1000$ of size $N_{TR} = 11$ is built by random sampling with replacement from D_{TR}. The sets $D_{TR,q}$ are then used to train $B = 1000$ different classification models $\{M_1, M_2, ..., M_{1000}\}$.

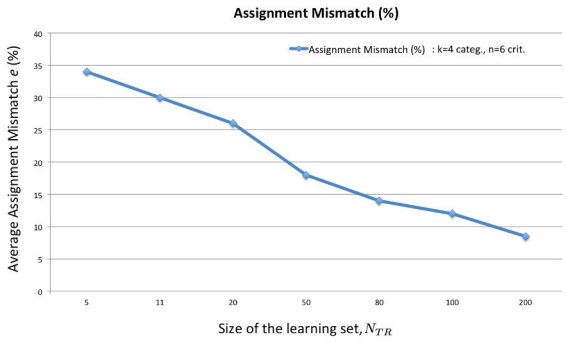

Fig. 5. Average Assignment error ϵ (%) as a function of the size N_{TR} of the learning set according to the model retrieval-based approach of §5.1.

This ensemble of models can be used to classify new alternatives. Figure 7 shows the probability distributions $P(A_h|x_p), h = 1, 2, ..., k = 4, p = 1, 2, ..., N_{TR} = 11$, empirically generated by the ensemble of $B = 1000$ bootstrapped MR-Sort classification models in the task of classifying the $N_{TR} = 11$ alternatives of the training set $D_{TR} = \{x_1, x_2, ..., x_{N_{TR}}\}$. The categories highlighted by the rectangles are those selected by the majority of the classifiers of the ensemble: It can be seen that the assigned classes coincide with the original categories of the alternatives of the training set (Table I), i.e., the accuracy of the inferred classification model based on the given training set (with 11 assigned alternatives) is 1.

In order to investigate the confidence of the algorithm in the classification of the test patterns, the results achieved

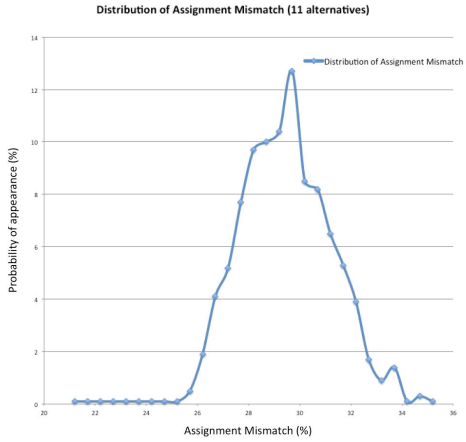

Fig. 6. Distribution of the assignment mismatch for a MR-Sort model trained with $N_{TR} = 11$ alternatives (%).

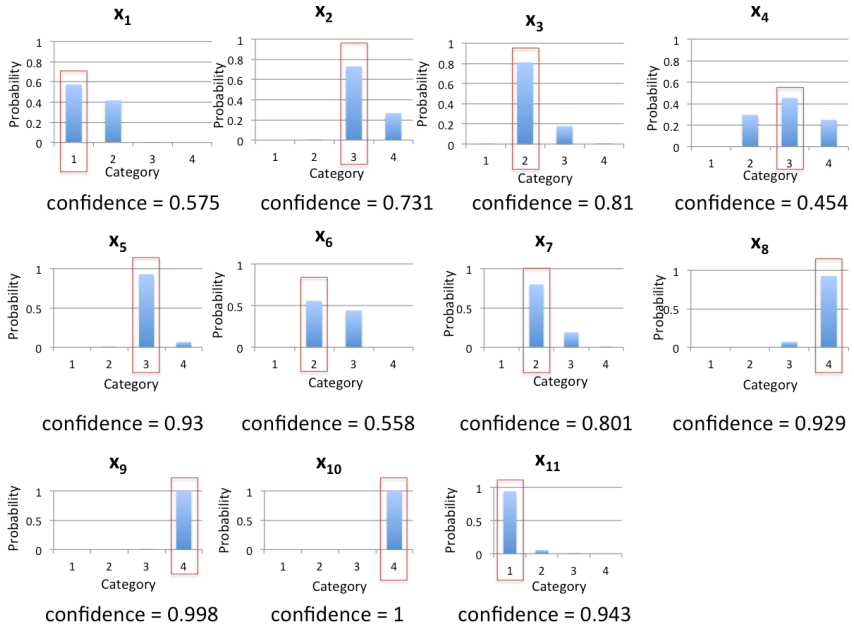

Fig. 7. Probability distributions $P(A_h|x_p), h = 1, 2, ..., k = 4, p = 1, 2, ..., N_{TR} = 11$ obtained by the ensemble of $B = 1000$ bootstrapped MR-Sort models in the classification of the alternatives x_p contained in the training set D_{TR}.

Table II . Number of patterns classified with confidence value

Confidence range	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]
$Numberofpatterns$	1	2	0
Confidence range	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]
$Numberofpatterns$	1	2	5

testing one specific pattern taken in turn from the training set are analysed. For each test of a specific pattern x_i, the distribution of the assignments by the $B = 1000$ classifiers shows the confidence of the assignment of the classification model on this specific pattern. By way of example, it can be seen that alternative x_3 is assigned to Class A^2(the correct one) with a confidence of $P(A^2|x_3) = 0.81$, whereas alternative x_6 is assigned to the same class A^2, but with a confidence of only $P(A^2|x_6) = 0.56$.

Notice that the most interesting information regards the confidence in the assignment of the test pattern to the class with the highest number of votes, i.e., the class actually assigned by the ensemble system according to the majority voting rule adopted [(10)]. In this respect, Table II reports the distribution of the confidence values associated to the class to which each of the 11 alternatives has been assigned.

Thus, a $10/11 \approx 91\%$ of all class assignments with confidence bigger than 0.5 are correct.

6.3 Application of the Leave-One-Out Cross-Validation (LOOCV) Method

Based on the original training set D_{TR} of size $N_{TR} = 11$, we generate 11 "new" training sets $D_{TR,i}, i = 1, 2, ..., 11$(each containing $N_{TR} - 1 = 1-$ assigned alternatives) by taking out each time one of the alternatives from D_{TR}. These 11 training sets are, then, used to train 11 different classification models $M_1, M_2, ..., M_{11}$. Each of these 11 models is used to classify the alternative correspondingly taken-out. Table III shows the comparison between the real classes Γ_p^t of the alternatives of the training set and the categories assigned by the trained models.

It can be seen that $\epsilon = 2$ out of the $N_{TR} = 11$ alternatives are assigned incorrectly (alternatives x_4 and x_6). Thus, the accuracy in the classification is given by the complement to 1 of the average error rate, i.e., $1 - \epsilon/N_{TR} = 1 - 2/11 = 1 - 0.182 = 0.818$. Notice that the 95% confidence interval for this recognition rate is [0.5901, 1].

7. DISCUSSION OF THE RESULTS

The three proposed methods provide conceptually and practically different estimates of the performance of the MR-Sort classification model.

The model retrieval-based approach provides a quite general indication of the classification capability of a vulnerability model with given characteristics. Actually, in this approach the only constant, fixed parameters are the size N_{TR} of the training set (given by the number of real-world classification examples available), the number of criteria n and the number of categories k (given by the analysts according to the characteristics of the systems at hand). On this basis, the space of all possible training sets of size N_{TR} and the space of all possible models with the above mentioned structure (n criteria and k categories) are randomly explored (again, notice that no use is made of the original real training set): the

Table III . Comparison between the real categories and the assignments provided by the LOOCV models

Alternative	Real Categories, Γ_p^t	Assignments by LOOCV method
x_1	1	1
x_2	3	3
x_3	2	2
x_4	3	2
x_5	3	3
x_6	2	3
x_7	2	2
x_8	4	4
x_9	4	4
x_{10}	4	4
x_{11}	1	1

classification performance is obtained as an average over the possible random training sets (of fixed size) and random models (of fixed structure). Thus, the resulting accuracy estimate is a realistic indicator of the expected classification performance of an 'average' model (of given structure) trained with an 'average' training set (of given size). In the case study considered, the average assignment error (resp. accuracy) is around 30% (resp. 70%).

On the contrary, the bootstrap method uses the real training set available to build an ensemble of models compatible with the data set itself. In this case, we do not explore the space of all possible training sets as in the model retrieval-based approach, but rather the space of all the classification models compatible with that particular training set constituted by real-world examples. In this view, the bootstrap approach serves the purpose of quantifying the uncertainty intrinsic in the particular (training) data set available when used to build a classification model of given structure (i.e., with given numbers n and k of criteria and categories, respectively). In this case study, the accuracy evaluated by the bootstrap method is much higher (equals to one) than that estimated by the model retrieval-based approach: this is reasonable because the latter evaluates the accuracy on a wider (i.e., in a broad sense, more uncertain) space of possible models and training sets; on the other hand, in the former method the training set adopted is given and it represents possibly only one of those randomly generated within the model retrieval-based approach. In addition, notice that differently from the model retrieval-based approach, the bootstrap method does not provide only the global classification performance of the vulnerability model, but also the confidence that for each test pattern a class assigned by the model is the correct one: this is given in terms of the full probability distribution of the vulnerability classes for each alternative to be classified.

Finally, also the leave-one-out cross-validation method has been used to quantify the expected classification performance of the model trained with the particular training data set available. In order to maximally exploit the information contained in the training set D_{TR}, $N_{TR} = 1$ "reduced" (training) sets are built, each containing $N_{TR} - 1 = 10$ assigned alternatives: each "reduced" set is used to build a model whose classification performance is evaluated on the element

correspondingly left out. The average error rate (resp. accuracy) turns out to be 18.2% (resp. 72.8%). The 95% confidence interval for the error rate (resp. accuracy) is approximately [0, 0.4099] (resp. [0.5901, 1]).

8. CONCLUSIONS

In this paper, the issue of quantifying the vulnerability of safety-critical systems (in the example, nuclear power plants) with respect to intentional hazards has been tackled within an empirical classification framework. To this aim an MR-Sort model has been trained by means of a small-sized set of data representing a priori-known classification examples. The performance of the MR-sort model has been evaluated with respect to: (i) its classification *accuracy*(resp., error), i.e., the expected fraction of patterns correctly (resp., incorrectly) classified; (ii) the *confidence* associated to the classification assignments (defined as the probability that the class assigned by the model to a given (single) pattern is the correct one). The performance of the empirically constructed classification model has been assessed by resorting to three approaches: a model retrieval-based approach, the bootstrap method and the leave-one-out cross-validation technique. To the best of the authors' knowledge, it is the first time that:

- a classification-based hierarchical framework is applied for the analysis of the vulnerability of safety-critical systems to intentional hazards;
- the confidence in the assignments provided by an MR-Sort classification model is quantitatively assessed by the bootstrap method in terms of the probability that a given alternative is correctly classified.

From the results obtained it can be concluded that although the model retrieval-based approach may be useful for providing an upper bound on the error rate of the classification model (obtained by exploring the space of all possible random models and training sets), the bootstrap method seems to be advisable for the following reasons: (i) it makes use of the training data set available from the particular case study at hand, thus characterising the uncertainty intrinsic in it; (ii) for each alternative (i.e., safety-critical system) to be classified, it is able to assess the confidence in the classification by providing the probability that the selected vulnerability class is the correct one. This is of paramount importance in the decision-making processes involving the vulnerability assessment of safety-critical systems, since it provides a metric for quantifying the 'robustness' of a given decision.

REFERENCES

1. Kröger W, Zio E. Vulnerable Systems. UK: Springer, 2001.
2. Aven T. Foundations of Risk Analysis. Wiley, N.J, 2003.
3. Aven T. Some reflections on uncertainty analysis and management. Reliability Engineering and System Safety, 2010; 95, 195-201.
4. Aven T. Misconceptions of Risk. Chichester: Wiley, 2010.
5. Aven T, Heide B. Reliability and validity of risk analysis. Reliability Engineering and System Safety, 2009; 94, 1862?1868.
6. Leroy A, Mousseau V, Pirlot M. Learning the Parameters of a Multiple Criteria Sorting Method Based on a Majority Rule, The Second International Conference on Algorithmic Decision Theory, 2011.
7. Aven T, Flage R. Use of decision criteria based on expected values to support decision-making in a production assurance and safety setting. Reliability Engineering and System Safety, 2009; 94, 1491-1498.

8. Milazzo MF, Aven T. An extended risk assessment approach for chemical plants applied to a study related to pipe ruptures. Reliability Engineering and System Safety, 2012; 99, 183-192.
9. Rocco C, Zio E. Bootstrap-based techiniques for computing confidence intervals in monte carlo system reliability evaluation. Reliability and Maintainability Symposium, 2005. Proceedings. Annual. Page(s): 303 ? 307.
10. Baraldi P, Razavi-Far R, Zio E. A method for estimating the confidence in the identification of nuclear transients by a bagged ensemble of FCM classifiers. NPIC&HMIT 2010.
11. Doumpos M., Zopounidis C., Multricriteria Decision Aid Classification Methods, Kluwer Academic Publishers, Netherlands. 2002, ISBN 1- 4020-0805-8.
12. NWRA, N. W. R. A. Risk assessment methods for water infrastructure systems, 2012.
13. Hofmann M, Kjølle G, Gjerde O. Development of indicators to monitor vulnerabilities in power systems, 2012.
14. Wang TR., Mousseau V, Zio E. A hierarchical decision making framework for vulnerability analysis, European Safety and RELiability Conference (ESREL 2013).
15. Roy B. The outranking approach and the foundations of ELECTRE methods. Theory and Decision 31, 1991, 49-73.
16. Mousseau V., Slowinski R. Inferring an ELECTRE TRI Model from Assignment Examples. Journal of Global Optimization, vol. 12, 1998, 157-174.
17. Efron B, Thibshirani RJ. An introduction to the bootstrap. Monographs on statistics and applied probability 57, 1993. Chapman and Hall, New York.
18. Zio E, A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. IEEE Transactions on Nuclear Science, 53(3), 2006; pp.1460-1470.
19. Cadini F, Zio E, Kopustinskas V, Urbonas R. An empirical model based bootstrapped neural networks for computing the maximum fuel cladding temperature in a RBMK-1500 nuclear reactor accident. Nuclear Engineering and Design, 238, 2008; pp. 2165-2172.
20. Baraldi P, Razavi-Fra R, Zio E. Bagged ensemble of fuzzy C means classifiers for Nuclear Transient Identification. Annals of Nuclear Energy 38, 5, 2010
21. Wilson R, Martinez TR. Combining cross-validation and confidence to measure fitness. Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), 1999; paper 163
22. Gutierrez-Osuna, R. Pattern analysis for machine olfaction: A review. IEEE SENSORS JOURNAL, 2002,10.1109/JSEN.2002.800688

APPENDIX A

As described in §2, the hierarchical model developed in [14] is considered to analyse the vulnerability of Nuclear Power Plants (NPPs) to intentional hazards. The susceptibility to intentional hazards (first layer) is characterised in terms of attractiveness and accessibility (second layer). These are hierarchically broken down into factors which influence them, including resilience seen as pre-attack protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness); this decomposition is made in 6 criteria: physical characteristics, social criticality, possibility of cascading failures, recovery means, human preparedness and level of protection (third layer). These six third-layer criteria are further decomposed into a layer of basic sub-criteria, for which data and information can be collected (fourth layer) (see Table I). The criteria of the layers are assigned preference directions for treatment in the decision-making process. The preference direction of a criterion indicates towards which state it is desirable to lead it to reduce susceptibility, i.e., it is assigned from the point of view of the defender of an attack who is concerned with protecting the system. Although only the 6 criteria of the third level of the hierarchy are considered in the NPPs vulnerability analysis considered in the present paper, examples of evaluation of the basic sub-criteria of the fourth

Table I . Criteria, sub-criteria and preference directions

Criterion	Physical characteristics	Social criticality	Possibility of cascading failures
Sub-criteria	Number of workers	Percentage of contribution to the welfare	Connection distance
	Nominal power production	Size of served cities	
	Number of production units		
Preference direction	Min	Min	Min
Criterion	Recovery means	Human preparedness	Level of protection
Sub-criteria	Number of installed backup components	Training	Physical size of the system
	Duration of backup components	Safety management	Number of accesses
	Duration of repair and recovery actions		Entrance control
	External emergency measures		Surveillance
Preference direction	Max	Max	Max

Table II . Number of workers

Level	Number of workers
1	500
2	1000
3	1500
4	2000
5	2500

layer are proposed in what follows for exemplification purposes: in particular, we describe an example of the procedure employed to calculate the numerical values of the third layer criteria on the basis of the characteristics of the fourth layer sub-criteria.

In extreme synthesis, the sub-criteria of the fourth layer can be characterised by crisp numbers or linguistic terms, depending on the nature of the sub-criterion. These descriptive terms and/or values of the fourth-layer sub-criteria are then scaled into numerical categories. The influence to the corresponding third-layer criterion of each of the sub-criteria is analysed.

To get the values of the six main third-layer criteria, (i) we assign arbitrary weights to each sub-criterion and (ii) we apply a simple weighted sum to the categorical values of the constituent sub-criteria.

A.1 Illustrative example: evaluation of the criterion ?Physical characteristics?

The criterion “Physical characteristics” is taken as an illustrative example. It is constituted by the sub-criteria “number of workers”, “nominal power production” and “number of production” or “service units”. The description and category scales are presented as follows:

Table III . Nominal power production

Level	Nominal power production
1	1000 MWe
2	3000 MWe
3	5000 MWe
4	7000 MWe
5	10000 MWe

Table IV . Number of production or service units

Level	Number of production or service units
1	2
2	4
3	6

Number of workers

This criterion can be seen to contribute to the attractiveness for an attack from various points of view, for example: 1) the more workers, the more work injuries and deaths from an attack; 2) the more workers, the easier for the attackers to sneak into the system; 3) the more workers, the higher the possibility that one of them can be turned into an attacker. Limiting the number of workers can, then, contribute to the security of the plant and, thus, reduce its attractiveness for an attack. Table II reports some reference values, typical of NPPs.

Nominal capacity

The higher the production capacity, the larger the potential consequences of lost production or security in case of an attack. Then, it is preferable to have a site with low capacity. Of course, for a fixed amount of total capacity needed, this would lead to its distribution on multiple sites, with an increase in the number of multiple targets, though each of them would lead to milder consequences if attacked. Table III shows some reference values of power generation capacity at NPP sites.

Number of production or service units

Locally, within a single site, this criterion represents the number of potential attack points. Preference would go towards having a small number of targets on a site. Table IV gives some reference values for NPPs.

We choose nuclear power plant x_1 as an example to show the calculation of the numerical value associated to the main criterion "Physical characteristics" starting from the data relative to the three corresponding sub-criteria (i.e., number of workers, nominal power production and number of production or service units). The original data of the three sub-criteria of x_1 is listed in Table V.

In scaling them onto corresponding category, we obtain the categorical value of alternative x_1(Table VI).

Table V . Corresponding sub-criteria original data of main criterion Physical characteristics of x_1

Alternative	Number of workers	Nominal power production (MWe)	Number of production or service units
x_1	600	1000	2

Table VI . Categorical value for the sub-criteria corresponding to the main criterion "Physical characteristics" of nuclear power plant x_1

Alternative	Number of workers	Nominal power production	Number of production or service units
x_1	2	2	1

Table VII . Normalised categorical value for corresponding sub-criteria of main criterion Physical characteristics of x_1

Alternative	Number of workers	Nominal power production	Number of production or service units
x_1	0.4	0.4	0.33

Table VIII . Weights of sub-criteria for Physical characteristics

Main Criterion: Physical Characteristics	Number of workers	Nominal power production	Number of production or service units
weights	0.3	0.5	0.2

Then, the numerical values of Table VI are normalised (that is, rescaled between 0 and 1 based on the pre-defined scales) as shown in Table VII.

Using the weights of these three sub-criteria (arbitrarily assigned by the authors) in Table VIII, we can apply a simple weighted sum to calculate the cumulative value for main criterion "Physical characteristics": $0.4*0.3+0.4*0.5+0.33*0.2 = 0.386$.

Finally, considering the preference directions of Table I (i.e., minimisation for criterion "Physical characteristics") and setting for each main criteria the value "0" as the worst case and "1" as the best one, we convert the cumulative weighed value obtained above to its complement to "1", i.e., $1 - 0.386 = 0.614$.

For the other five main third-layer criteria, the process of calculation is the same as for criterion "Physical characteristics".

APPENDIX B

Mathematical details about the algorithm of disaggregation of a MR-Sort classification model

We consider the case involving k categories that are, thus, separated by *(k-1)* frontier denoted $b = \{b^1, b^2, ..., b^h, ..., b^{k-1}\}$, where $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h, h = 1, 2, ..., k\}$, n is the number of criteria that are taken into account. Let $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{TR}\}$ be the training set, where N_{TR} is the number of alternatives and, $(A^1, A^2, ..., A^k)$ be the partition of the training set, ordered from the ?best? to ?worst? alternatives.

For each alternative $x_p \in D_{TR}$, in category A^h of the learning set D_{TR}(for $h = 2, 3, ..., k-1$), let us define 2n binary variables δ_{ip}^h and δ_{ip}^{h-1}, for $p = 1, 2, ..., N_{TR}$, such that δ_{ip}^l equals to 1 iff $g_i(x_p) \geq b_i^l$ for $l = h-1, h$ and $\delta_{ip}^h = 0 \Leftrightarrow g_i(x_p) < b_i^h$. We introduce 2n continuous variables $c_{ip}^l (l = h-1, h)$ constrained to be equal to ω_i if $\delta_{ip}^l = 1$ and to *0* otherwise.

We consider an objective function that describes the robustness of the assignment. We introduce two more continuous variables, y_p and z_p, for each $x_p \in D_{TR}$ and α. In maximising α, we maximise the value of the minimal slack in the constraints.

We resume all the constraints in the following mathematical program:

$$max\ \alpha, \tag{A.1}$$

$$\alpha \leq y_p, \alpha \leq z_p, \forall x_p \in D_{TR}, \tag{A.2}$$

$$\sum_{i,p \in \mathbb{N}} c_{ip}^l + y_p + \epsilon = \lambda, \forall x_p \in A^{l-1}, \tag{A.3}$$

$$\sum_{i,p \in \mathbb{N}} c_{ip}^l = \lambda + z_p, \forall x_p \in A^l, \tag{A.4}$$

$$c_{ip}^l \leq \omega_i, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A.5}$$

$$c_{ip}^l \leq \delta_{ip}^l, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A.6}$$

$$c_{ip}^l \geq \delta_{ip}^l - 1 + \omega_i, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A.7}$$

$$M\delta_{ip}^l + \epsilon \geq g_i(x_p) - b_i^l, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A.8}$$

$$M(\delta_{ip}^l - 1) \leq g_i(x_p) - b_i^l, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A.9}$$

$$\sum_{i,p \in \mathbb{N}} \omega_i = 1, \lambda \in [0.5, 1], \tag{A.10}$$

$$\omega_i \in [0, 1], \forall i \in \mathbb{N}, \tag{A.11}$$

$$c_{ip}^l \in [0, 1], \delta_{ip}^l \in \{0, 1\}, \forall x_p \in D_{TR}, \forall i \in \mathbb{N}, \tag{A.12}$$

$$y_p, z_p \in \mathbb{R}, \forall x_p \in D_{TR}, \tag{A.13}$$

$$\alpha \in \mathbb{R}, \tag{A.14}$$

M is an arbitrary large positive value, and ϵ, arbitrary small positive quantity.

The case in which x_p belongs to one of the extreme categories (A^1 and A^k) is simple. It requires the introduction of only n binary variables and n continuous variables. In fact, if x_p belongs to A^1 we just have to express that the

subset of criteria on which x_p is at least as good as b_1 has sufficient weight. In a dual way, when x_p lies in A^k, the worst category, we have to express that it is at least as good as b_k on a subset of criteria that has not sufficient weight.

Paper (iii)

An empirical classification-based framework for the safety-related criticality assessment of complex energy production systems, in presence of inconsistent data

T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio.

Reliability Engineering & System Safety, 2015, under review.

An empirical classification-based framework for the safety-related criticality assessment of complex energy production systems, in presence of inconsistent data

Tai-Ran WANG[a], Vincent MOUSSEAU[b], Nicola PEDRONI[a], Enrico ZIO[a,c]

[a]*Chair on Systems Science and the Energy challenge, Fondation EDF, Ecole Centrale Paris and Supelec*

[b]*Laboratory of Industrial Engineering, Ecole Centrale Paris, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex*

[c]*Politecnico di Milano, Energy Department, Nuclear Section, c/o Cesnef, via Ponzio 33/A , 20133, Milan, Italy, Fax: 39-02-2399.6309, Phone: 39-02-2399.6340, enrico.zio@polimi.it*

ABSTRACT

The technical problem addressed by the present paper is the assessment of the safety-related criticality of complex energy production systems. An empirical classification model is developed, based on the Majority Rule Sorting method, to perform the evaluation of the class of criticality of the plant/system of interest. The model is built on the basis of a (limited-size) set of data representing (a priori-known) criticality classification examples provided by experts.

The empirical construction of the classification model may raise two issues. First, the classification examples provided may contain contradictions: a validation of the consistency of the considered data set is, thus, required. Second, uncertainty affects the evaluation process: a quantitative assessment of the performance of the classification model is, thus, in order, in terms of accuracy and confidence in the assignments.

In this paper, two approaches are used to tackle the first issue: the inconsistencies in the data examples are "resolved" by deleting or relaxing, respectively, some constraints in the process of model construction. Three methods are considered to address the second issue: (i) a model retrieval-based approach, (ii) the Bootstrap method and (iii) the cross-validation technique.

Numerical analyses are presented with reference to a case study involving Nuclear Power Plants.

KEYWORDS: Safety-criticality, classification model, data consistency validation, confidence estimation, MR-Sort, nuclear power plants

1. INTRODUCTION

The ever-growing attention to Energy and Environmental (E&E) issues has led to emphasize a systemic view involving the trilemma of energy systems' safety and security, sustainable development and cost effectiveness [(1)]. In particular, the assessment of the level of safety-related criticality of the existing complex energy production systems for, possibly, their "informed" enhancement is extremely demanded. This has sparked a number of efforts to guide designers, managers and stakeholders in (i) the definition of the criteria for the evaluation of safety criticality, (ii) its qualitative and quantitative assessment [(2)(3)] and (iii) the selection of actions to reduce its level of criticality. In this paper, we mainly address the central issue (ii) above, i.e., the quantitative assessment of the level of safety-related criticality of complex, energy production systems, with particular reference to Nuclear Power Plants (NPPs).

The analysis of the safety-related criticality of a system generally involves many different sources of uncertainty [(4)] such as long time frame, capital intensive investment and the involvement of a multiple stakeholders with different views and preferences [(5)(6)]. Thus, it is difficult to proceed with traditional risk/safety assessment methods, such as statistical analysis or probabilistic modeling [(7)].

In this paper, motivated by EDF (Electricité de France), as a methodology for aiding their decisions on selection of alternative safety barriers, maintenance options etc., we propose to tackle the problem within an empirical classification framework, to develop a classification model based on the Majority Rule Sorting (MR-Sort) method [(10)] (which is a simplified ELECTRE model [(8)(9)]) to assign an alternative of interest (i.e., a complex system or plant) to a given (criticality) class (or category). The MR-Sort classification model contains a group of (adjustable) parameters that have to be calibrated by means of a set of *empirical* classification examples (also called training set), i.e., a set of alternatives with the corresponding (criticality) classes pre-assigned by *experts*.

Such empirical construction of the classification model may raise two practical issues. First, the classification examples provided by the experts may contain contradictions: a validation of the consistency of the data set is, thus, required. In this paper, two approaches are used to tackle this issue: the

inconsistencies in the data examples are "resolved" by *deleting* or *relaxing*, respectively, some constraints in the process of model construction [10]. Second, due to the finite (typically small) size of the set of training classification examples usually available for the analysis of real complex safety-critical systems, the performance of the classification model is impaired. In particular, (i) the classification *accuracy* (resp., error), i.e., the expected fraction of patterns correctly (resp., incorrectly) classified, is typically reduced (resp., increased); (ii) the classification process is characterized by significant uncertainty, which affects the *confidence* of the classification-based evaluation model. In our work, we define the confidence in a classification assignment as in Ref. 10, i.e., as the probability that the class assigned by the model to a given (single) pattern is the correct one. The performance of the classification model (i.e., the classification accuracy – resp., error – and the confidence in the classification) needs to be quantified: this is of paramount importance for taking robust decisions informed by the integrated evaluation of the level of safety criticality of complex energy production systems [11][12]. In this paper, three different approaches are used to assess the performance of a classification-based MR-Sort evaluation model in the presence of small training data sets. The first is a model-retrieval based approach [10], which is used to assess the expected percentage error in assigning new alternatives. The second is Cross-Validation (CV): a given number of alternatives from the entire database is randomly selected to form the training set and generate the corresponding model, which is, then, used to classify the rest of the alternatives. By so doing, the expected percentage model error is estimated as the fraction of alternatives incorrectly assigned (as an average over the left-out data). The third, is based on *bootstrapping* the available training set in order to build an ensemble of evaluation models [13]; the method can be used to assess both the accuracy and the confidence of the model: in particular, the confidence in the assignment of a given alternative is given in terms of the full (probability) distribution of the possible performance classes for that alternative (built on the bootstrapped ensemble of evaluation models) [14].

The contribution of this work is threefold:

- a verification of the consistency of the classification examples provided by the experts is carried out by resorting to two approaches: as a result, the training data set is modified accordingly before the process of model construction;
- classification models are used in a variety of fields including finance, marketing, environmental and energy management, human resources management, medicine, risk analysis, fault diagnosis etc. [15]: to the best of the authors' knowledge, this is the first time that a classification-based framework is applied for the evaluation of the safety-related criticality of complex energy production systems (e.g., Nuclear Power Plants);
- to the best of the authors' knowledge, it is the first time that the confidence in the assignments provided by an MR-Sort classification model is quantitatively assessed by the bootstrap method, in terms of the probability that a given alternative is correctly classified.

The paper is organized as follows. The next Section presents the basic framework for system criticality evaluation. Section 3 shows the classification model applied within the proposed framework. Section 4 describes the learning process of a classification model by the disaggregation method. Section 5 deals with the inconsistency study of the pre-assigned data set. In Section 6, three approaches are proposed to analyze the performance of the classification model. Then, the proposed approaches are applied in Section 7 to a case study involving a set of nuclear power plants. Finally, Sections 8 and 9 present the discussion of the results and the conclusions of this research, respectively.

2 GENERAL FRAMEWORK FOR THE EVALUATION OF THE LEVEL OF SAFETY-RELATED CRITICALITY OF COMPLEX ENERGY PRODUCTION SYSTEMS AND PLANTS

Without loss of generality, we consider that the overall level of criticality of the system is characterized in terms of a set of six criteria $x' = \{x_1', x_2', x_3', x_4', x_5', x_6'\}$: its level of safety, its level of security and protection, its possible impact on the environment, its long-term performance, its operational performance

and its possible impact on the communication and image of the operational enterprise (Figure 1.). These six criteria are used as the basis to assess the level of criticality of the system. Each criterion is evaluated in 4 grades, ranging from best (grade '0') to worst (grade '3'). Further details about the "scoring" of the criticality of each criterion are given in Appendix A. Four levels (or categories) of criticality are considered: satisfactory (0), acceptable (1), problematic (2) and serious (3). Then, the assessment of the level of criticality can be performed within a classification framework: find the criticality category (or class) corresponding to the evaluation of the system in terms of the six criteria above. A description of the algorithm used to this purpose is given in the following Section.

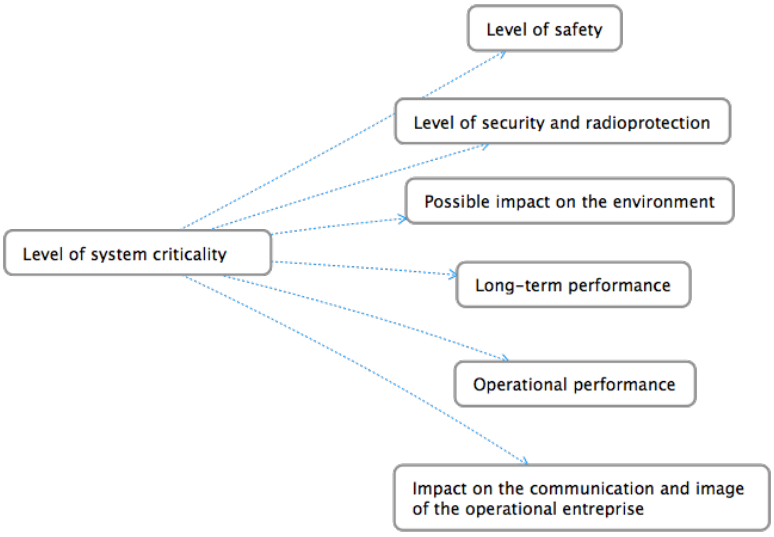

Figure 1. Criteria used to characterize the overall level of criticality of a complex energy production system or plant.

3 CLASSIFICATION MODEL FOR THE EVALUATION OF THE LEVEL OF CRITICALITY OF COMPLEX ENERGY PRODUCTION SYSTMES: THE MAJORITY RULE SORTING (MR-SORT) METHOD

The Majority Rule Sorting Model (MR-Sort) method is a simplified version of ELECTRE Tri, an outranking sorting procedure in which the assignment of an alternative to a given category is determined

using a complex concordance non-discordance rule [8][9]. We assume that the alternative to be classified (in this paper, a safety-critical energy production system, e.g., a nuclear power plant) can be described by an *n*-tuple of elements $x = \{x_1, x_2, ..., x_i, ..., x_n\}$ (in our case, a 6-tuple x as described in the previous section), which represent the evaluation of the alternative with respect to a set of *n* criteria (by way of example, in the present paper the 6 criteria used to evaluate the level of safety-related criticality of the system of interest include safety, security, impact on the environment and so on, as described in Section 2). We denote the set of criteria by $nCrit = \{1, 2, ..., i, ..., n\}$ and assume that the values x_i of criterion *i* range in the set X_i [12]. The MR-Sort procedure allows assigning any alternative $x = \{x_1, x_2, ..., x_i, ..., x_n\} \in X = X_1 \times X_2 \times ... \times X_i \times ... \times X_n$ to a particular pre-defined category (in this paper, a class of overall criticality), in a given ordered set of categories, $\{A^h : h = 1, 2, ..., k\}$; as mentioned in Section 2, k = 4 categories are considered in this work: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, A^4 = serious.

To this aim, the model is further specialized in the following way:

-We assume that X_i is a subset of $\mathcal{R}$ for all $i \in N$ and the sub-intervals $(X_i^1, X_i^2, ..., X_i^h, ..., X_i^k)$ of X_i are compatible with the order on the real numbers, i.e., for all $x_i^1 \in X_i^1, x_i^2 \in X_i^2, ..., x_i^h \in X_i^h, ..., x_i^k \in X_i^k$, we have $x_i^1 > x_i^2 > ... > x_i^h > ... > x_i^k$. For example, in the present paper all the criteria are ranged from best (grade '0') to worst (grade '3'). Regarding the preference of the algorithm (bigger values are preferred) and the normalization purpose for the final value which combines all the characteristics of the chosen criteria, the original data are preferred to be transformed into new values within a range of [0, 1] by simple basic calculations: $\{x | x_i = 1 - x_i'/3, i \in \{1,2, \dots ,6\}\}$. We assume, furthermore, that each interval $X_i^h, h = 2, 3, ..., k$ has a smallest element b_i^h, which implies that $x_i^{h-1} \geq b_i^h > x_i^h$. The vector $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}$ (containing the lower bounds of the intervals X_i^h of criteria $i = 1, 2, ..., n$ in correspondence of category *h*) represents the lower limit profile of category A^h.

-There is a weight ω_i associated with each criterion $i = 1, 2, ..., n$, quantifying the relative importance of criterion *i* in the evaluation assessment process; notice that the weights are

normalized such that $\sum_{i=1}^{n} \omega_i = 1$.

In this framework, a given alternative $x = \{x_1, x_2, ...x_i, ..., x_n\}$ can be assigned to a range of [0,1], with "0" representing the worst situation, and "1" the best one. One alternative is assigned to category $A^h, h = 1, 2, ..., k$, iff

$$\sum_{i \in N: x_i \geq b_i^h} \omega_i \geq \lambda \quad \text{and} \quad \sum_{i \in N: x_i \geq b_i^{h+1}} \omega_i < \lambda, \qquad (1)$$

where λ is a threshold $(0 \leq \lambda \leq 1)$ chosen by the analyst. Rule *(1)* is interpreted as follows. An alternative x belongs to category A^h if: (1) its evaluations in correspondence of the *n* criteria (i.e., the values $\{x_1, x_2, ..., x_i, ..., x_n\}$) are at least as good as b_i^h (lower limit of category A^h with respect to criterion *i*), $i = 1, 2, ..., n$, on a subset of criteria that has sufficient importance (in other words, on a subset of criteria that has a weight larger than or equal to the threshold λ chosen by the analyst); and at the same time (2) the weight of the subset of criteria on which the evaluations $\{x_1, x_2, ..., x_i, ..., x_n\}$ are at least as good as b_i^{h+1} (lower limit of the successive category A^{h+1} with respect to criterion *i*), $i = 1, 2, ..., n$, is not sufficient to justify the assignment of x to the successive category A^{h+1}.

Notice that alternative x is assigned to the best category A_1 if $\sum_{i \in N: x_i \geq b_i^1} \omega_i \geq \lambda$ and it is assigned to the worst category A_k if $\sum_{i \in N: x_i \geq b_i^{k-1}} \omega_i < \lambda$. Finally, it is straightforward to notice that the parameters of such a model are the $(k-1) * n$ lower limit profiles (*n* limits for the *k-1* categories, since the worst category doesn't need one), the *n* weights of the criteria $\omega_1, \omega_2, ..., \omega_i, ..., \omega_n$, and the threshold λ, for a total of $k * n + 1$ parameters.

4 CONSTRUCTING THE MR-SORT CLASSIFICATION MODEL

In order to construct an MR-Sort classification model, we need to determine the set of $k * n + 1$ parameters described in the previous Sections, i.e., the weights $\omega = \{\omega_1, \omega_2, ..., \omega_n\}$, the lower profiles $b = \{b^1, b^2, ..., b^h, ..., b^k\}$, with $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}, h = 1, 2, ..., k$, and the threshold λ; in this paper, λ is considered a

fixed, constant value chosen by the analyst (e.g., λ=0.9).

To this aim, the decision maker provides a training set of "classification examples" $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{TR}\}$, i.e., a set of N_{TR} alternatives (in this case, nuclear power plants) $x_p = (x_1^p, x_2^p, ...x_i^p, ..., x_n^p)$, $p = 1, 2, ..., N_{TR}$ together with the corresponding real pre-assigned categories (i.e., criticality classes) Γ_p^t (the superscript 't' indicates that Γ_p^t represents the true, a priori-known performance class of alternative x_p).

The calibration of the $k * n$ parameters is done through the learning process detailed in [6]. In extreme synthesis, the information contained in the training set D_{TR} is used to restrict the set of MR-Sort models compatible with such information, and to finally select one among them [6]. The a priori-known assignments generate constraints on the parameters of the MR-Sort model. In [6], such constraints have a linear formulation and are integrated into a Mixed Integer Program (MIP) that is designed to select one (optimal) set of such parameters ω^* and b^* (in other words, to select one classification model $M(\cdot \mid \omega^*, b^*)$) that is coherent with the data available and maximizes a defined *objective function*. In [6], the optimal parameters $\omega^*$and b^* are those that maximize the value of the minimal slack in the constraints generated by the given set of data D_{TR}. Once the (optimal) classification model $M(\cdot \mid \omega^*, b^*)$ is constructed, it can be used to assign a new alternative x (i.e., a new nuclear power plant) to one of the performance classes $A^h, h = 1, 2, ..., k$: in other words, $M(x \mid \omega^*, b^*) = \Gamma_x^M$ where Γ_x^M is the class assigned by model $M(\cdot \mid \omega^*, b^*)$ to alternative x and assumes one value among $\{A^h : h = 1, 2, ..., k\}$. Further mathematical details about the training algorithm are not given here for brevity: the reader is referred to [6] and to Appendix B at the end of the paper.

There are two main issues related to this disaggregation process and to the empirical construction by the MR-Sort classification model. First, for the given set of pre-assigned alternatives, it is possible that some of the assignments are not consistent, due to fact that different experts may give different judgments upon similar situation (which causes an internal inconsistency); thus, the given data set has to be made consistent in order to obtain a compatible empirical classification model. Second, because of the finite and quite small number N_{TR} of available classification examples in most of real applications involving the

evaluation of safety-critical systems, the model $M(\cdot \mid \omega^*, b^*)$ is only a partial representation of reality and its assignments are affected by uncertainty: this uncertainty needs to be quantified to build confidence in the decision process, which follows the criticality level assessment.

In the following Section, the methods used in this paper to study the consistency of a given training dataset are described in detail; then, in Section 6 three different methods are presented to assess the performance of the MR-sort classification model.

5 CONSISTENCY STUDY: VALIDATION AND MODIFICATION OF THE SET OF ALTERNATIVES PRE-ASSIGNED BY EXPERTS

As highlighted before, sorting models consist in assigning alternatives evaluated on several criteria to ordered categories. To implement such models, it is necessary to set the values of the preference parameters used in the model. Rather than fixing the values of these parameters directly, a usual approach is to infer these values from assignment examples provided by experts and decision makers (DMs). However, assignment examples provided by experts and DMs can be *inconsistent*, i.e., may not "produce" any meaningful classification model. Such a situation can be understood according to two perspectives: either the examples provided by the DM contradict each other, or the preference model is not flexible enough to account for the way the DM assigns alternatives holistically. In the first case, the DM would acknowledge a misjudgment and would agree to reconsider his/her examples; in the second case, the DM would not agree to change the examples and the preference model should be changed. In both cases, we refer to an inconsistency situation. In any case, the DM needs to know what causes inconsistency, i.e., which judgments should be changed if the aggregation model is to be kept (which is our case)[16].

The MIP algorithm summarized in the previous section may prove infeasible in case the assignments of the alternatives in the learning set are incompatible with all MR-sort models. In order to help the DMs to understand how their inputs are conflicting and to question previously expressed judgments, to learn about their preferences as the interactive process evolves, we formulate two MIPs that are able to: (i) find one

MR-sort model that maximize the number of learning set alternatives correctly assigned and (ii) propose accordingly a possible modification for each of the conflicting alternatives.

5.1 Inconsistency resolution via constraints deletion

Resolving the inconsistencies can be performed by deleting a subset of constraints related to the inconsistent alternatives. As shown in Figure 2, each alternative x_p can provide one or two constraints with respect to its assignment: for example, alternatives assigned to extreme categories, i.e., A_1 and A_4, provide one constraint, whereas alternatives assigned to intermediate categories, i.e., A_2 and A_3, introduce two constraints. Let us introduce a binary variable γ_p for each alternative x_p, which is equal to "1" if *all* the constraints associated to x_p are fulfilled, and equal to "0" otherwise.

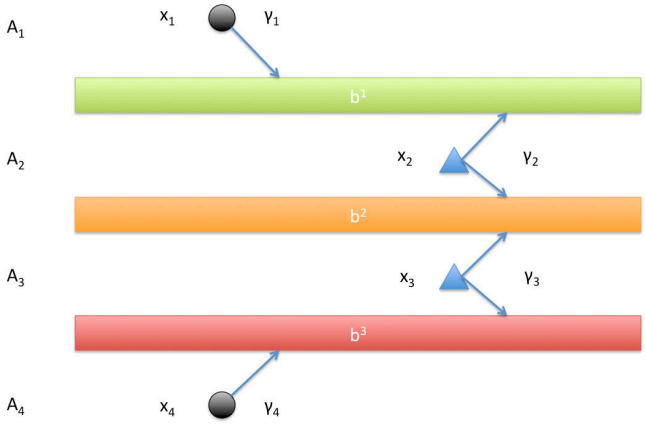

Figure 2. Representation of constraints deletion algorithm

The algorithm proceeds by "deleting" (i.e., removing) those constraints (i.e., those alternatives) that do not allow the creation of a compatible classification model, while maximizing the number of alternatives retained in the learning set (i.e., in minimizing the number of alternatives that are not taken into account): by so doing, we maximize the quantity of information that can be used to generate a classification model correctly. In other words, we obtain a MIP that yields a subset $D^*_{TR} \subseteq D_{TR}$ of maximal cardinality that can be represented by an MR-sort model. The reader is referred to Appendix B at the end of the paper for more mathematical details.

5.2 Inconsistency resolution via constraints relaxation

Based on the algorithm presented in the previous subsection, a subset of maximal cardinality that can be represented by an MR-sort model is obtained. At the same time, its complementary set is *deleted*. However, in order to help the DMs understand in what way the identified inconsistent inputs conflict with the others; and guide them to reconsider and possibly modify their judgments, a constraints relaxation algorithm is here proposed.

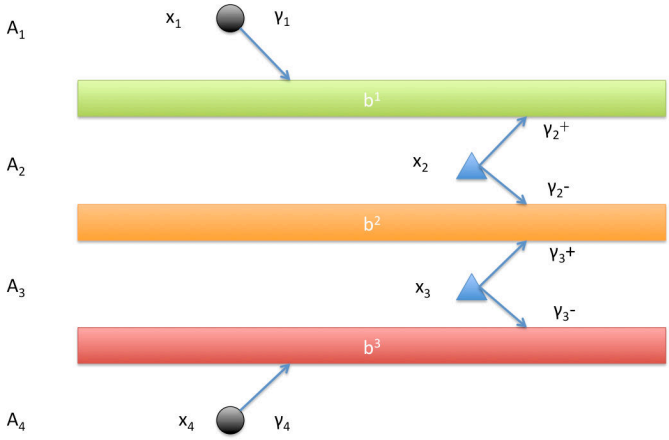

Figure 3. Representation of constraints relaxation algorithm

As presented in Section 5.3, each alternative x_p can provide one or two constraints with respect to its assignment. As presented in Figure 3, we introduce the following binary variables: γ_p, for the alternatives originally assigned to extreme categories, i.e., A_1 and A_4; γ_p^+ and γ_p^- for the alternatives originally assigned to intermediate categories, i.e., A_2 and A_3: In particular, γ_p^+ refers to the fulfillment of the constraint associated to the better category lower profiles, whereas γ_p^- refers to the fulfillment of the constraint associated to the worse category lower profiles.

As in the previous case, the algorithm identifies a subset $D_{TR}^* \subseteq D_{TR}$ of maximal cardinality that can generate an MR-sort model with proper formulation. In addition, for each of the alternatives that are not accepted into the subset D_{TR}^*, the corresponding inconsistent constraints are also targeted: for example, if

for one alternative x_p we obtain $\gamma_p^+ = 0$ (resp., $\gamma_p^- = 0$), then this alternative should be classified in a better (resp., worse) category; in other words, its original assignment is underestimated (resp., overestimated). The same criterion is applied to the alternatives that are originally assigned to the best or worst category. The reader is referred to Appendix B at the end of the paper for more mathematical details.

6 METHODS FOR ASSESSING THE PERFORMANCE OF THE CLASSIFICATION-BASED MODEL FOR THE EVALUATION OF THE LEVEL OF CRITICALITY OF COMPLEX ENERGY PRODUCTION SYSTEMS AND PLANTS

6.1 Model Retrieval-Based Approach

The first method of performance assessment is based on the model-retrieval approach proposed in [6]. A fictitious set D_{TR}^{rand} of N_{TR} alternatives $\{x_p^{rand} : p = 1, 2, ..., N_{TR}\}$ is generated by random sampling within the ranges X_i of the criteria, $i = 1, 2, ..., n$. Notice that the size N_{TR} of the fictitious set D_{TR}^{rand} has to be the same as the real training set D_{TR} available, for the comparison to be fair. Also, a MR-Sort classification model $M(\cdot \mid \omega^{rand}, b^{rand})$ is constructed by randomly sampling possible values of the internal parameters, $\{\omega_i : i = 1, 2, ..., n\}$ and $\{b_h : h = 1, 2, ..., k-1\}$. Then, we simulate the behavior of a Decision Maker (DM) by letting the (random) model $M(\cdot \mid \omega^{rand}, b^{rand})$ assign the (randomly generated) alternatives $\{x_p^{rand} : p = 1, 2, ..., N_{TR}\}$. In other words, we construct a learning set D_{TR}^{rand} by assigning the (randomly generated) alternatives using the (randomly generated) MR-Sort model, i.e., $D_{TR}^{rand} = \{(x_p^{rand}, \Gamma_p^M) : p = 1, 2, ..., N_{TR}\}$, where Γ_p^M is the class assigned by model $M(\cdot \mid \omega^{rand}, b^{rand})$ to alternative x_p^{rand}, i.e., $\Gamma_p^M = M(x_p^{rand} \mid \omega^{rand}, b^{rand})$. Subsequently, a new MR-Sort model $M'(\cdot \mid \omega', b')$, compatible with the training set D_{TR}^{rand}, is inferred using the MIP formulation summarized in Section 3 and in the Appendix B. Although models $M(\cdot \mid \omega^{rand}, b^{rand})$ and $M'(\cdot \mid \omega', b')$ may be quite different, they coincide on the way they assign elements of D_{TR}^{rand}, by construction. In order to compare models M and M′, we randomly generate a (typically large) set D_{test}^{rand} of *new* alternatives $D_{test}^{rand} = \{x_p^{test,rand} : p = 1, 2, ..., N_{Test}\}$ and we compute the percentage of "assignment errors", i.e., the proportion of these N_{Test} alternatives that models

M and M′ assign to different criticality categories.

In order to account for the randomness in the generation of the training set D_{TR}^{rand} and of the model $M(\cdot \mid \omega^{rand}, b^{rand})$, and to provide robust estimates for the assignment errors ε, the procedure outlined above is repeated for a large number N_{sets} of random training sets $D_{TR}^{rand,j}, j = 1, 2, ..., N_{sets}$; in addition, for each set j the procedure is repeated for different random models $M(\cdot \mid \omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models}$. The sequence of assignment errors thereby generated, $e_{jl}, j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$, is, then, averaged to obtain a robust estimate for ε. The procedure is sketched in Figure 4.

Notice that this method does not make any use of the original training set D_{TR} (i.e., of the training set constituted by real-world classification examples). In this view, the model retrieval-based approach can be interpreted as a tool to obtain an absolute evaluation of the expected error that an 'average' MR-Sort classification model $M(\cdot \mid \omega, b)$ with k categories, n criteria and trained by means of an 'average' data set of given size N_{TR} makes in the task of classifying a new generic (unknown) alternative.

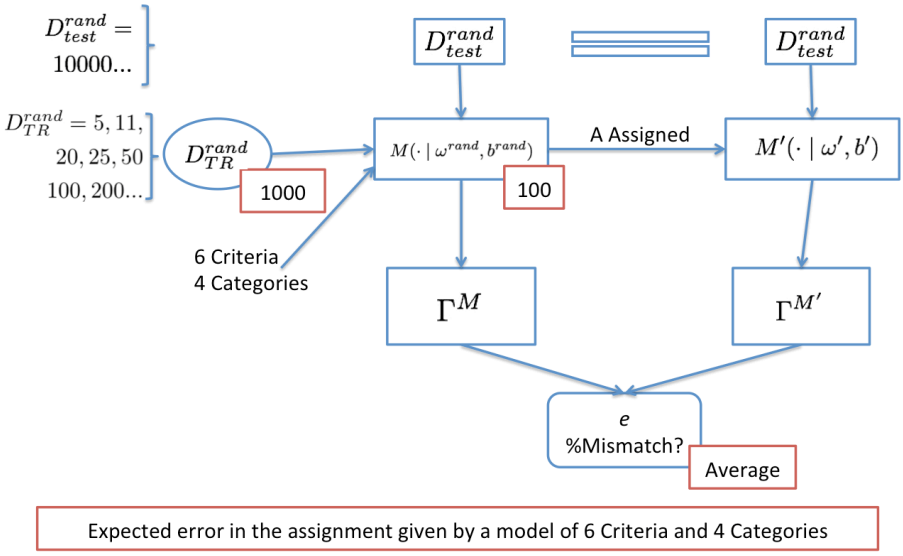

Figure 4. The general structure of the model-retrieval approach

6.2 Cross-Validation Technique [17][18][19]

This approach is proposed to characterize the performance of the MR-Sort model in terms of average classification accuracy (resp., error).

The method proceeds as follows:

0. Set the iteration number q=1;

1. For a data set $D = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{total}\}$ with pre-assigned alternatives, select a learning set $D_{TR}^q = \{(x_s, \Gamma_s^t), s = 1, 2, ..., N_{TR}\}$ (with $\frac{1}{2} N_{Total} < N_{TR} < N_{Total}$) by performing random sampling without replacement from the given D. The remaining alternatives are used to form a test set $D_{TS}^q = \{(x_r, \Gamma_r^t), s = 1, 2, ..., N_{TS}\}$, with $N_{TS} = N_{Total} - N_{TR}$.

2. Build a classification model $\{M_q(\cdot|\omega_q, b_q)\}$ on the basis of the training set $D_{TR} = \{(x_s, \Gamma_s^t), s = 1, 2, ..., N_{TR}\}$.

3. Use the classification model $\{M_q(\cdot|\omega_q, b_q)\}$ to provide a class Γ_r^q to the elements of the corresponding test set $D_{TS}^q = \{(x_r, \Gamma_r^t), s = 1, 2, ..., N_{TS}\}$.

4. The classification error ϵ^q on test set D_{TS}^q is computed as the fraction of alternatives of D_{TS}^q that are incorrectly classified.

Steps 1-4 are repeated for $q = 1, 2, ..., B$ times (in this paper, *B*=1000). Finally, the expected classification error of the algorithm is obtained as the average of the classification errors $\epsilon^q, q = 1, 2, ..., B$, obtained on the *B* test sets D_{TS}^q, $q = 1, 2, ..., B$. The general structure of the algorithm is as shown in Figure 5.

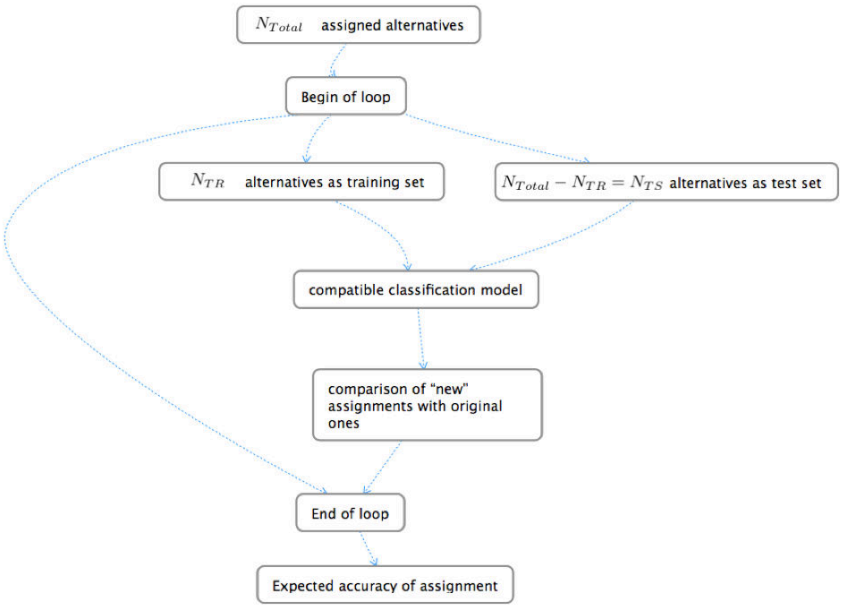

Figure 5. The general structure of the Cross-Validation Technique

6.3 The Bootstrap Method

A way to assess both the accuracy (i.e., the expected fraction of alternatives correctly classified) and the confidence of the classification model (i.e., the probability that the category assigned to a given alternative is the correct one) is by resorting to the bootstrap method [20], which is used to create an ensemble of classification models constructed on different data sets bootstrapped from the original one [21]. The final class assignment provided by the ensemble is based on the combination of the individual output of classes provided by the ensemble of models [13].

The basic idea is to generate different training datasets by random sampling with replacement from the original one [22]. The different training sets are used to build different individual classifications. The individual classifiers of the ensemble perform well possibly in different regions of the training space and, thus, they are expected to make errors on alternatives with different characteristics; these errors are balanced out in the combination, so that the performance of the ensemble is, in general, superior to that of the single classifiers [21][22].

In this paper, the output classes of the single classifiers are combined by *majority voting*: the class chosen by most classifiers is the ensemble final assignment. The bootstrap-based empirical distribution of the assignments given by the different classification models of the ensemble is used to measure the confidence in the classification of a given alternative x, that represents the probability that such alternative is correctly assigned [13][22].

In more details, the main steps of the bootstrap algorithm here developed are as follows (Figure 6):

1. Build an ensemble of B (typically of the order of 500-1000) classification models $\{M_q(\cdot \mid (\omega_q, b_q) : q = 1, 2, ..., B\}$ by random sampling with replacement from the original data set D_{TR} and use each of the bootstrapped models $M_q(\cdot \mid \omega_q, b_q)$ to assign a class Γ_x^q, q = 1, 2,..., B, to a given alternative x of interest (notice that Γ_x^q takes a value in $A_h, h = 1, 2, ..., k$). By so doing, a bootstrap-based empirical probability distribution $P(A_h \mid x), h = 1, 2, ..., k$ for category A_h of alternative x is produced, which is the basis for assessing the confidence in the assignment of alternative x. In particular, repeat the following steps for q = 1, 2,..., B:

a. Generate a bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$, by performing random sampling with replacement from the original data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$ of N_{TR} input/output patterns. The data set $D_{TR,q}$ is, thus, constituted by the same number N_{TR} of input/output patterns drawn among those in D_{TR}, although due to the sampling with replacement some of the patterns in D_{TR} will appear more than once in $D_{TR,q}$, whereas some will not appear at all.

b. Build a classification model $\{M_q(\cdot \mid \omega_q, b_q) : q = 1, 2, ..., B\}$, on the basis of the bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{TR}\}$.

c. Use the classification model $M_q(\cdot \mid \omega_q, b_q)$ to provide a class $\Gamma_x^q, q = 1, 2, ..., B$ to a given alternative of interest, i.e., $\Gamma_x^q = M_q(x \mid \omega_q, b_q)$.

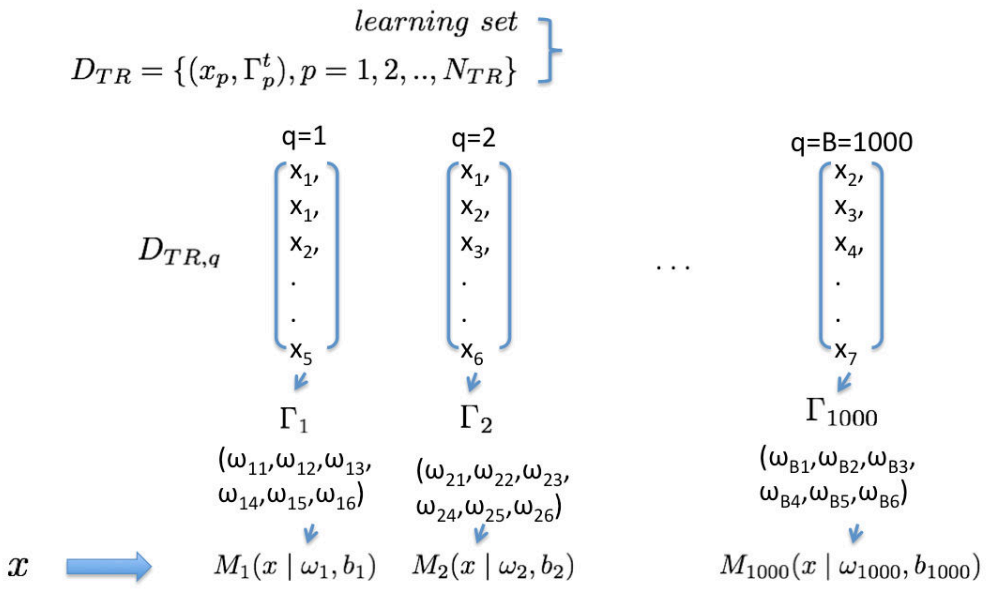

Figure 6. The bootstrap algorithm

2. Combine the output classes $\Gamma^q, q = 1, 2, ..., B$ of the individual classifiers by majority voting: the class chosen by most classifiers is the ensemble assignment Γ_x^{ens}, i.e., $\Gamma_x^{ens} = argmax_{A^h}[card_q\{\Gamma_x^q = A^h\}]$.

3. As an estimation of the confidence in the majority-voting assignment Γ_x^{ens} (step 2, above), we consider the bootstrap-based empirical probability distribution $P(A_h \mid x), h = 1, 2, ..., k$, i.e., the probability that category A_h is the correct category given that the (test) alternative is x [6]. The

estimator of $P(A_h \mid x)$ here employed is: $P(A_h \mid x) = \frac{\sum_{q=1}^{B} I\{\Gamma_q = A_h\}}{B}$, where $I\{\Gamma_q = A_h\} = 1$, if $\Gamma_q = A_h$, and *0* otherwise.

4. Finally, the accuracy of classification is presented by the estimator $P(A_h \mid x)$ (ratio of the number of alternatives correctly assigned by the classification models to the total number of alternatives). The error of the classification model is defined as the complement to 1 to the accuracy.

7 APPLICATIONS

The methods presented in Sections 4 - 6 are applied on an exemplificative case study concerning the assessment of the overall level of safety-related criticality of Nuclear Power Plants (NPPs) [9]. We identify n = 6 main criteria $i = 1, 2, ..., n = 6$ by means of the approach presented in [9] (see Section 2): x_1 = level of safety, x_2 = level of security and radioprotection, x_3 = possible impact on the environment, x_4 = long-term performance, x_5 = operational performance and x_6 = impact on the communication and image of the enterprise. Then, k = 4 criticality categories $A_h, h = 1, ..., k = 4$ are defined as: A_1 = satisfactory, A_2 = acceptable, A_3 = problematic and A_4 = dangerous (Section 2). The entire original dataset is constituted by a group of 35 systems x_p with the corresponding a priori-known category Γ_p^t (Table I).

In what follows, first we apply the two approaches for data consistency validation (Section 7.1); then, we use the three techniques of Section 6 to assess the performance of the MR-Sort classification-based model built using the training set D_{TR} (Section 7.2).

Table I. Original training data set

	Criticality evaluation criteria						
Alternatives (NPPs)	Safety	Secrurity and Radioprotection	Impact on the Environment	Long-term Performance	Operational Performance	Communication and Image of the Operational Enterprise	Category Assignment (original set)
x1	3	0	3	3	0	2	3
x2	1	0	1	1	0	2	1
x3	1	0	1	2	0	1	2
x4	2	2	3	0	0	1	2
x5	3	1	2	3	0	1	3
x6	1	3	2	2	0	1	2
x7	2	0	3	2	0	3	4
x8	2	2	3	2	0	0	1
x9	1	0	2	0	0	0	1
x10	2	0	3	0	0	2	3
x11	2	0	3	2	0	2	3
x12	1	0	3	1	0	1	3
x13	1	0	2	0	0	1	1
x14	2	0	0	0	0	1	2
x15	1	0	0	0	0	0	1
x16	1	0	0	0	0	1	3
x17	2	0	0	2	0	1	3
x18	1	2	2	0	0	1	2
x19	0	0	0	0	0	1	3
x20	0	3	0	0	1	0	4
x21	1	0	2	1	1	0	4
x22	1	3	0	0	1	0	2
x23	1	0	1	0	1	0	1
x24	1	0	2	0	0	0	4
x25	1	0	0	0	1	0	1
x26	1	0	0	0	0	0	2
x27	1	0	0	0	0	1	2
x28	1	0	0	0	0	1	2
x29	2	2	3	0	0	0	3
x30	2	2	3	2	0	0	2
x31	2	2	2	1	0	0	1
x32	3	0	3	0	0	3	2
x33	1	0	1	0	0	0	3
x34	3	0	0	1	0	3	3
x35	3	0	0	0	0	3	2

7.1 Consistency study results

The application of the MR-sort disaggregation algorithm on the given set of alternatives $D = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N = 35\}$ (Table I) does not lead to the generation of any classification model (infeasible solution by the MIP algorithm), because there are inconsistencies within the given data. There may exist different types of inconsistencies, as illustrated in Table II by two examples:

Table II. Examples of inconsistent assignments

Case 1:

	Criticality evaluation criteria						
Alternatives (NPPs)	Safety	Secrurity and Radioprotection	Impact on the Environment	Long-term Performance	Operational Performance	Communication and Image of the Operational Enterprise	Category Assignment (original set)
x16	1	0	0	0	0	1	3
x27	1	0	0	0	0	1	2

Case 2:

	Criticality evaluation criteria						
Alternatives (NPPs)	Safety	Secrurity and Radioprotection	Impact on the Environment	Long-term Performance	Operational Performance	Communication and Image of the Operational Enterprise	Category Assignment (original set)
x13	1	0	2	0	0	1	1
x19	0	0	0	0	0	1	3

In Case 1, two alternatives (x_{16} and x_{27}) with same value for all the six criteria are assigned to different categories (resp., 3 and 2). In Case 2, an alternative (x_{19}) with better characteristics than another (x_{13}) with respect to the six criteria, is assigned to a worse category (3).

Such inconsistencies are solved via constraints deletion (Section 7.1.1) and constraints relaxation (Section 7.1.2).

7.1.1 Inconsistency resolution via constraints deletion

We first consider finding out the consistent dataset with maximized number of pre-assigned alternatives. We analyze the given data set by the constraints deletion algorithm. In the given set D of 35 alternatives, 14 are deleted, which leaves a consistent data set of 21 alternatives. The new consistent set $D_{ad} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{ad} = 21\}$ is, then, used to generate a compatible classification model $\{M_{ad}(\cdot|(\omega_{ad}, b_{ad})\}$ by the MR-sort disaggregation algorithm. Then, all the alternatives in the original data set D are assigned a class by model M_{ad}: such assignments agree with the results of the constraints deletion process, i.e., only the deleted alternatives are not correctly assigned (see Table III, where the deleted alternatives are highlighted).

7.1.2 Inconsistency resolution via constraints relaxation

In the previous Section, we succeeded in obtaining a consistent data set from a given inconsistent one by deleting the inconsistent alternatives of a "wrong" assignment. However, from the point of view of the decision makers, it would be ideal to retain as many alternatives as possible in the training set, especially when the size of the ensemble is limited (which is always the case of the evaluation problem of safety-critical infrastructures). This can be done by modifying the pre-defined (wrong) assignments of the

inconsistent alternatives.

We examine the same set D by means of the constraints relaxation algorithm presented in section 5.2. After the application of the algorithm, we obtain the set $D_{ar} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{ar} = 21\}$, which is identical to the set $D_{ad} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N_{ad} = 21\}$ obtained in the previous subsection (for the alternatives in this set, the corresponding generated constraints are consistent). The remaining alternatives form the set $D_r = D - Dar$. However, this algorithm also allows the identification of two more sets: (i) $D_{up} = \{(x_p, \Gamma_p^t) | \gamma_p^+ = 0\}$ (i.e., the set of alternatives whose assignments should be better than the original one, indicated in Table III by a "+" in the shadowed Table cells in column "Constraint relaxation"); (ii) $D_{down} = \{(x_p, \Gamma_p^t) | \gamma_p^- = 0\}$ (i.e., the set of alternatives whose assignments should be worse than the original one, indicated in Table III by a "-" in the shadowed Table cells in the column "Constraints relaxation").

Based on the indications given by the sets D_{up} and D_{down}, we have modified each of the alternatives in D_r by one category in the direction suggested by the relaxation algorithm. Combining the alternatives thereby modified in D_r with the ones in D_{ar}, we obtain a new data set of 35 alternatives $D_{relax} = \{(x_p, \Gamma_p^{relax}) : p = 1, 2, ..., N_{relax} = 35\}$. A group of $N_{TR} = 25$ data of D_{relax} (marked as *"TR"* in the first column of Table III) is used to build the training set D_{TR} for the model, i.e., $D_{TR} = \{(x_p, \Gamma_p^{relax} : p = 1, 2, ..., N_{TR} = 25)\}$; the remaining 10 alternatives (marked as *"TS"* in the first column of Table III) are used for testing the model generated. In what follows, we consider the classification model generated using dataset D_{relax} and we assess its performance in terms of accuracy and confidence in the assignments.

Table III. Original inconsistent dataset and the corresponding modifications operated by the constraint deletion and relaxation algorithms

	Criticality evaluation criteria								
Alternatives (NPPs)	Safety	Secrurity and Radioprotection	Impact on the Environment	Long-term Performance	Operational Performance	Communication and Image of the Operational Enterprise	Category Assignment (original set)	Constraints deletion	Constraints relaxation
x1 (TR)	3	0	3	3	0	2	3	3	3
x2 (TR)	1	0	1	1	0	2	1	2	-
x3 (TR)	1	0	1	2	0	1	2	2	2
x4 (TR)	2	2	3	0	0	1	2	3	-
x5 (TR)	3	1	2	3	0	1	3	3	3
x6 (TR)	1	3	2	2	0	1	2	2	2
x7 (TR)	2	0	3	2	0	3	4	4	4
x8 (TR)	2	2	3	2	0	0	1	3	-
x9 (TR)	1	0	2	0	0	0	1	1	1
x10 (TR)	2	0	3	0	0	2	3	3	3
x11 (TR)	2	0	3	2	0	2	3	3	3
x12 (TR)	1	0	3	1	0	1	3	3	3
x13 (TR)	1	0	2	0	0	1	1	2	-
x14 (TR)	2	0	0	0	0	1	2	2	2
x15 (TR)	1	0	0	0	0	0	1	1	1
x16 (TR)	1	0	0	0	0	1	3	2	+
x17 (TR)	2	0	0	2	0	1	3	3	3
x18 (TR)	1	2	2	0	0	1	2	2	2
x19 (TR)	0	0	0	0	0	1	3	1	+
x20 (TR)	0	3	0	0	1	0	4	1	+
x21 (TR)	1	0	2	1	1	0	4	1	+
x22 (TR)	1	3	0	0	1	0	2	2	2
x23 (TR)	1	0	1	0	1	0	1	1	1
x24 (TR)	1	0	2	0	0	0	4	1	+
x25 (TR)	1	0	0	0	1	0	1	1	1
x26 (TS)	1	0	0	0	0	0	2	1	+
x27 (TS)	1	0	0	0	0	1	2	2	2
x28 (TS)	1	0	0	0	0	1	2	2	2
x29 (TS)	2	2	3	0	0	0	3	3	3
x30 (TS)	2	2	3	2	0	0	2	3	-
x31 (TS)	2	2	2	1	0	0	1	1	1
x32 (TS)	3	0	3	0	0	3	2	3	-
x33 (TS)	1	0	1	0	0	0	3	1	+
x34 (TS)	3	0	0	1	0	3	3	3	3
x35 (TS)	3	0	0	0	0	3	2	3	-

7.2 Methods for assessing the performance of the classification-based model for the evaluation of the criticality of complex systems

7.2.1 Application of the Model Retrieval-Based Approach

We generate $N_{sets} = 1000$ different training sets $D_{TR}^{rand,j}, j = 1, 2, ..., N_{sets}$, and for each set j, we randomly generate $N_{models} = 100$ models $M(\cdot \mid \omega^{rand,l}, b^{rand,l}), l = 1, 2, ..., N_{models} = 100$. By so doing, the expected accuracy (1-ε) of the corresponding MR-Sort model is obtained as the average of $N_{sets} \cdot N_{models} = 1000 \cdot 100 = 100000$ values $(1 - e_{jl}), j = 1, 2, ..., N_{sets}, l = 1, 2, ..., N_{models}$ (see Section 6.1). The size N_{test} of the random test set D_{TR}^{rand} is $N_{test} = 10000$. Finally, we perform the procedure of Section 6.1 for different sizes N_{TR} of the random training set D_{TR}^{rand} (even if the chosen size of the training set in our following case study is $N_{TR} = 25$, see Section 7.1.2): in particular, we choose $N_{TR} = 5, 11, 20, 25, 50, 100 \; and \; 200$. This analysis serves the purpose of outlining the behavior of the accuracy (1-ε) as a function of the amount of classification examples available.

The results are summarized in Figure 7, where the average percentage assignment error ε is shown as a function of the size N_{TR} of the learning set (from 5 to 200). As expected, the assignment error ε tends to decrease when the size of the learning set N_{TR} increases: the higher the cardinality of the learning set, the higher (resp. lower) the accuracy (resp. the expected error) in the corresponding assignments. Comparing these results with those obtained by Leroy et al [6] using MR-Sort models with $k = 2$ and 3 categories and n = 3-5 criteria, it can be seen that for a given size of the learning set, the error rate (resp. the accuracy) grows (resp. decreases) with the number of model parameters to be determined by the training algorithm = $k * n + 1$. It can be seen that for our model with $n = 6$ criteria and $k = 4$ categories, in order to guarantee an error rate inferior to 10% we would need training sets consisting of more than N_{TR}= 100 alternatives. Typically, for a learning set of N_{TR} = 25 alternatives (as chosen in Section 7.1.2), the average assignment error ε is around 24%; correspondingly, the accuracy of the MR-Sort classification model trained with the data set D_{TR} of size $N_{TR} = 25$ available in the present case is around (1-ε) = 76%: in other words, there is a probability of 76% that a new alternative (i.e., a new NPP) is assigned to the correct category of performance.

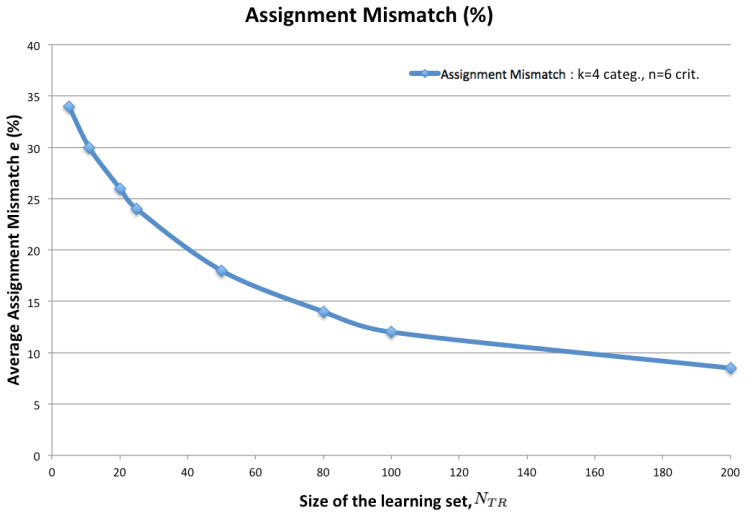

Figure 7. Average Assignment error ε (%) as a function of the size N_{TR} of the learning set according to the model retrieval-based approach of Section 5.1

In order to assess the randomness intrinsic in the procedure used to obtain the accuracy estimate above, we have also calculated the 95% confidence intervals for the average assignment error ε of the models trained with N_{TR} = 11, 20, 25 and 100 alternatives in the training set. The 95% confidence interval for the error associated to the models trained with 11, 20, 25 and 100 alternatives as learning set are [25.4%, 33%], [22.2%, 29.3%], [12.8%, 27.6%] and [10%, 15.5%], respectively. For illustration purposes, Figure 8 shows the distribution of the assignment mismatch built using the $N_{sets} \cdot N_{models} = 100000$ values $e_{jl}, j = 1, 2, ..., N_{sets} = 1000, l = 1, 2, ..., N_{models} = 100$, generated as described in Section 5.1 for the example of 25 alternatives.

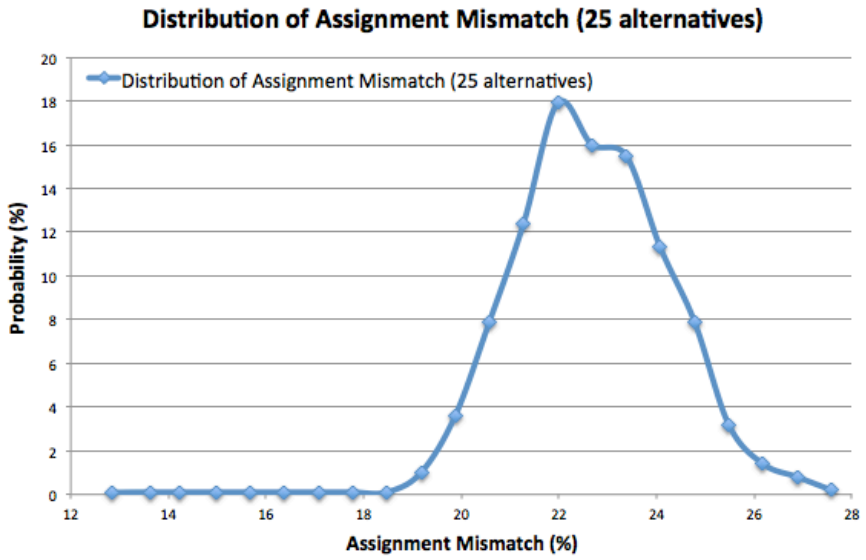

Figure 8. Distribution of the assignment mismatch for a MR-Sort model trained with N_{TR} = 25 alternatives (%)

7.2.2 Application of the Cross-Validation Technique

A loop of *B* (=1000) iterations is performed, as presented in section 6.2. We take D_{relax} as the training set and generate a learning set $D_{TR} = \{(x_p, \Gamma_p^{relax}) : p = 1, 2, ..., N_{TR} = 25\}$ for each loop by performing random sampling without replacement from it. The test set is formed by the corresponding complimentary set of D_{TR}. The average error calculated is around 18%.

7.2.3 Application of the Bootstrap Method

A number B (= 1000) of bootstrapped training sets $D_{TR,q}, q = 1, 2, ..., 1000$ of size N_{TR} = 25 is built by random sampling with replacement from D_{TR} (see Section 7.1.2). The sets $D_{TR,q}$ are then used to train B = 1000 different classification models $\{M_1, M_2, ..., M_{1000}\}$. Then, all the data available (both the training and test elements) are classified by the ensemble.

Notice that all the training patterns are assigned by majority voting to the correct class [13]: in other words, the accuracy of the ensemble of models on the training set is 100%. Then, a confidence in the assignment is also provided. In this respect, Table IV reports the distribution of the confidence values associated to the class to which each of the 25 alternatives has been assigned.

Table IV: Number of patterns classified with a given confidence value

Confidence range	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Number of patterns	1	3	1	11	9

Thus, a fraction of $20/25 \cong 80\%$ of all the alternatives (i.e., the critical plants) of the training set are correctly assigned with confidence bigger than 0.8.

The ensemble of models can also be used to classify new alternatives, e.g., the alternatives in the test set D_{TS} (see Section 7.1.2). Figure 9 shows the probability distributions of the 10 elements of $P(A_h|x_p), h = 1, 2, ..., k-4, p = 1, 2, ..., N_{TS} = 10$, empirically generated by the ensemble of B = 1000 bootstrapped MR-Sort classification models in the task of classifying the N_{TS} = 10 alternatives of the test set $D_{TS} = \{x_1, x_2, ..., x_{N_{TS}}\}$. The categories highlighted by the rectangles are the correct ones, as obtained by the constraints relaxation algorithm (Section 7.1.2, Table III). It can be seen that six alternatives (x_{26}, x_{27}, x_{28}, x_{29}, x_{30} and x_{33}) over 10 are correctly assigned: in other words, the accuracy of the informed bootstrapped ensemble is around $6/10 \cong 60\%$.

Then, for each specific test pattern x_i, the distribution of the assignments by the B = 1000 classifiers is analyzed to obtain the corresponding confidence. By way of example, it can be seen that alternative x_{28} is assigned to Class A^2 (the correct one) with a confidence of $P(A^2|x_{28}) = 0.931$, whereas alternative x_{26} is assigned to Class A^1 but with a confidence of only $P(A^1|x_{26}) = 0.856$.

More importantly, it can be seen that the 4 alternatives incorrectly classified (x_{31}, x_{32}, x_{34} and x_{35}) are assigned a class close to the correct one; in addition, the "true" class is given the second highest confidence in the distribution. For example, alternative x_{35} is assigned to class A^4 instead of A^3 with 68% confidence; however, the true Class A^3 is still given a confidence of 32%.

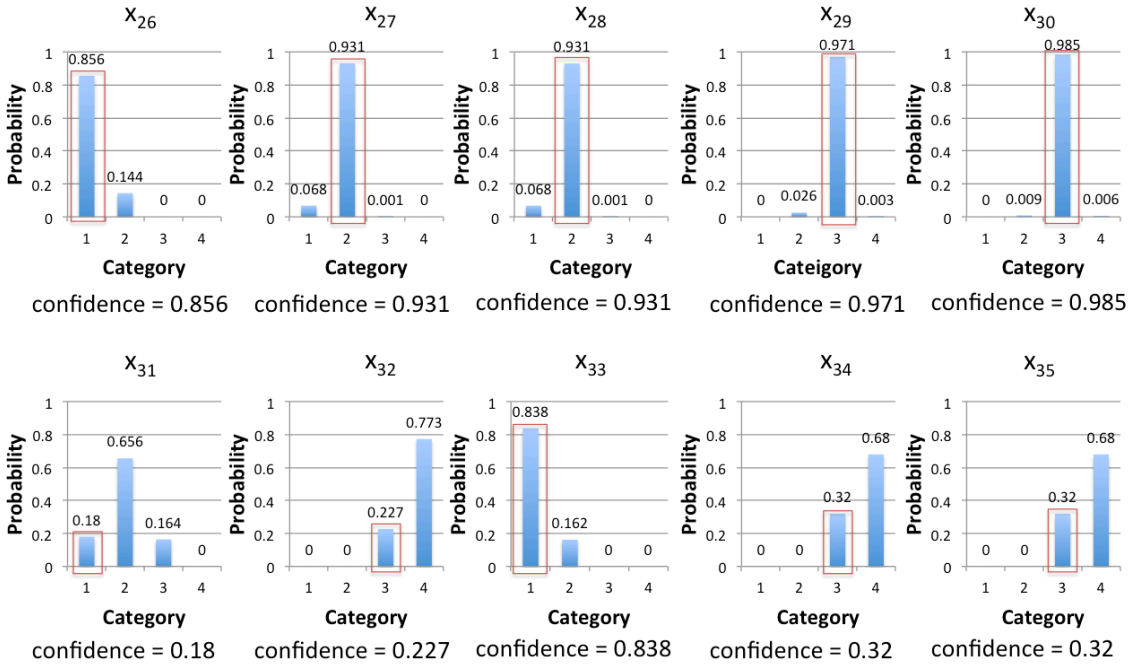

Figure 9. Probability distributions examples of $P(A_h|x_p), h = 1, 2, ..., k-4, p = 1, 2, ..., N_{TS} = 10$ obtained by the ensemble of B = 1000 bootstrapped MR-Sort models in the classification of the alternatives x_p contained in the training set D_{TR}

8 DISCUSSION OF THE RESULTS

The analysis of the inconsistencies of the original dataset has ensured the generation of a coherent training set and, correspondingly of a compatible classification model for system criticality evaluation: $D_{TR} = \{(x_p, \Gamma_p^{relax} : p = 1, 2, ..., N_{TR} = 25)\}$, generated by constraints relaxation.

Then, three methods have been used to assess the performance of the classification model thereby generated: the three methods provide conceptually and practically different estimates of the performance of the MR-Sort classification model.

The model retrieval-based approach provides a quite general indication of the classification capability of

an evaluation model with given characteristics. Actually, in this approach the only constant, fixed parameters are the size N_{TR} of the training set (given by the number of real-world classification examples available), the number of criteria *n* and the number of categories *k* (given by the analysts according to the characteristics of the systems at hand). On this basis, the space of all possible training sets of size N_{TR} and the space of all possible models with the above mentioned structure (*n* criteria and *k* categories) are randomly explored (again, notice that no use is made of the original real training set): the classification performance is obtained as an average over the possible random training sets (of fixed size) and random models (of fixed structure). Thus, the resulting accuracy estimate is a realistic indicator of the expected classification performance of an 'average' model (of given structure) trained with an 'average' training set (of given size). In the case study considered, the average assignment error (resp. accuracy) is around 24% (resp. 76%).

The cross-validation method has been also used to quantify the expected classification performance in terms of accuracy. In order to maximally exploit the information contained in the available data set, B=1000 training sets of size $N_{TR} = 25$ are generated by random sampling without replacement from the original set. Each training set is used to build a model whose classification performance is evaluated on the ten elements correspondingly left out. The average error rate (resp. accuracy) turns out to be 18% (resp. 82%).

On the contrary, the bootstrap method uses the training set available to build an ensemble of models compatible with the data set itself. In this case, we do not explore the space of all possible training sets as in the model retrieval-based approach, but rather the space of all the classification models compatible with that particular training set constituted by real-world examples. In this view, the bootstrap approach serves the purpose of quantifying the uncertainty intrinsic in the particular (training) data set available when used to build a classification model of given structure (i.e., with given numbers *n* and *k* of criteria and categories, respectively). In this case study, the accuracy evaluated by the bootstrap method is slightly lower than that estimated by the model retrieval-based approach, with an error (accuracy) rate equals 40% (60%). However, notice that differently from the model retrieval-based approach, the bootstrap method

does not provide only the global classification performance of the evaluation model, but also the confidence that for each test pattern a class assigned by the model is the correct one: this is given in terms of the full probability distribution of the performance classes for each alternative to be classified.

9 CONCLUSIONS

In this paper, the issue of assessing the criticality of complex energy production systems (in the example, nuclear power plants) with respect to different safety-related criteria has been tackled within an empirical framework of classification. An MR-Sort model has been trained by means of a small-sized set of data representing a priori-known criticality classification examples provided by experts. Inconsistencies and contradictions in the initial data set have been resolved by resorting to constraint deletion and relaxation algorithms that have maximized the number of consistent examples in the training set. The performance of the MR-sort model has been evaluated with respect to: (i) its classification *accuracy* (resp., error), i.e., the expected fraction of patterns correctly (resp., incorrectly) classified; (ii) the *confidence* associated to the classification assignments (defined as the probability that the class assigned by the model to a given (single) pattern is the correct one). In particular, the performance of the empirically constructed classification model has been assessed by resorting to three approaches: a model retrieval-based approach, the cross-validation technique and the bootstrap method. To the best of the authors' knowledge, it is the first time that:

- a classification-based framework is applied for the criticality assessment of energy production systems (e.g., Nuclear Power Plants) from the point of view of safety-related criteria;
- the confidence in the assignments provided by the MR-Sort classification model developed is assessed by the bootstrap method in terms of the probability that a given alternative is correctly classified.

From the results obtained in the case study, it can be concluded that although the model retrieval-based approach may be useful for providing an upper bound on the error rate of the classification model

(obtained by exploring the space of all possible random models and training sets), for practical applications the bootstrap method seems to be advisable for the following reasons: (i) it makes use of the training data set available from the particular case study at hand, thus characterizing the uncertainty intrinsic in it; (ii) for each alternative (i.e., safety-critical system) to be classified, it is able to assess the confidence in its classification by providing the probability that the selected performance class is the correct one. This seems of paramount importance in the decision-making processes performed by the assessed safety-criticality, since it provides a metric for the 'robustness' of the decision. The methodology can be further applied to more systems, e.g. the NRC's Risk-Informed Regulatory Oversight Program, in which reactors are placed in different classes, which affects the amount of regulatory oversight performed.

Acknowledgements:

The authors are thankful to François Beaudouin and Dominique Vasseur of EDF R&D for providing input classification examples and guidance throughout the work.

APPENDIX A. Criticality levels associated to the criteria used for the integrated assessment of a system from the point of view of safety criteria (Section 2)

In what follows, the criticality "scores" associated to each classification criterion introduced in Section 2 are specified.

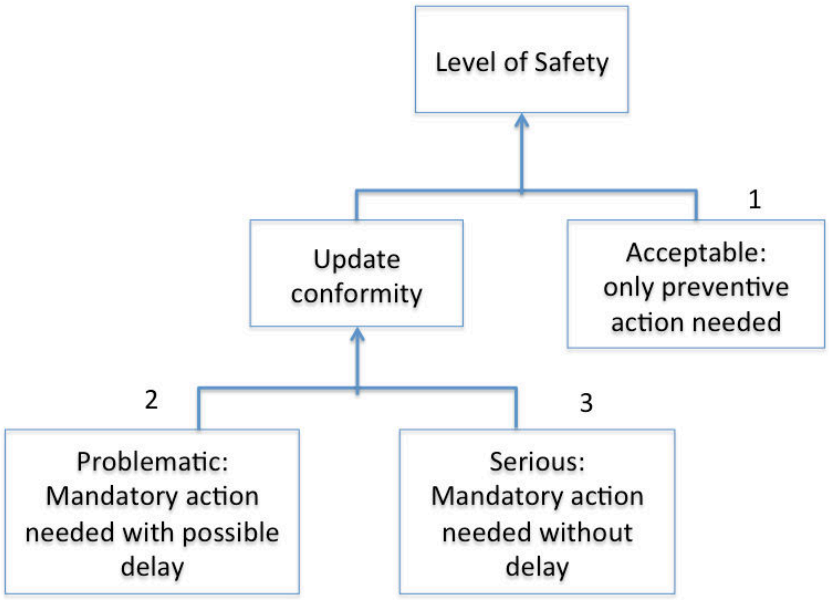

Figure A.1 "Scoring" of criticality for criterion "Level of Safety"

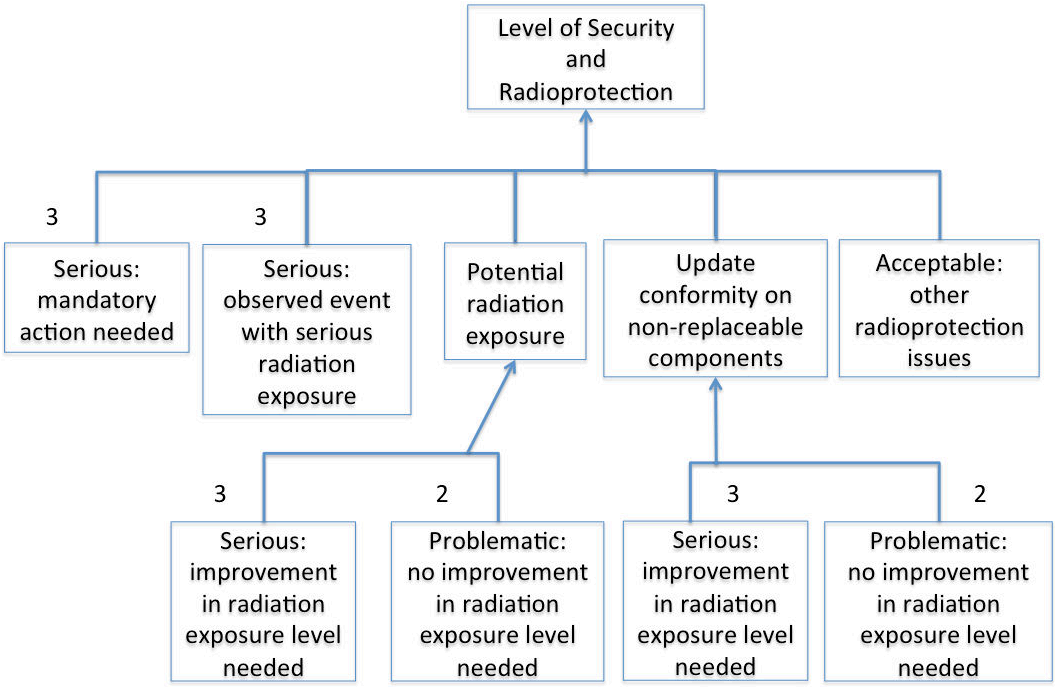

Figure A.2 “Scoring” of criticality for criterion “Level of Security and Radioprotection”

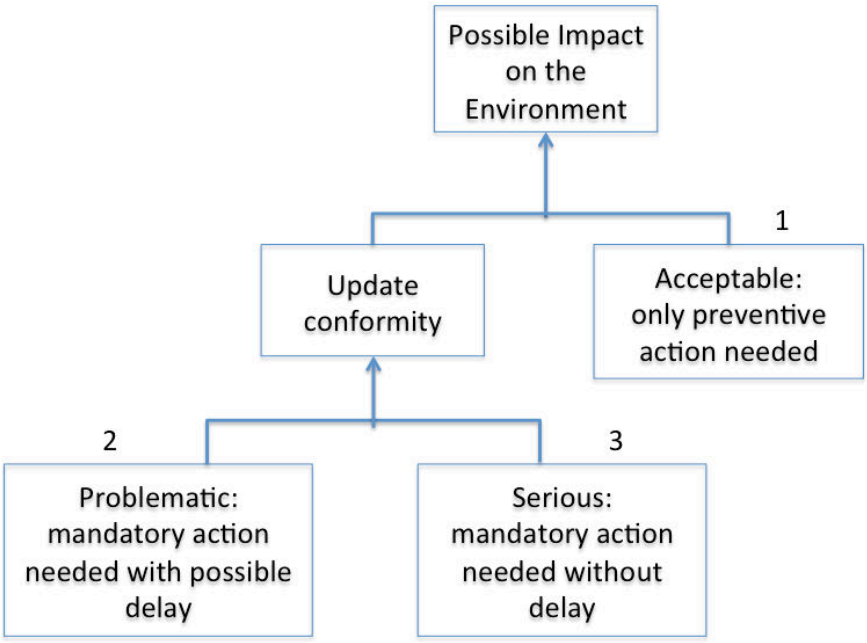

Figure A.3 “Scoring” of criticality for criterion “Level of Possible Impact on the Environment”

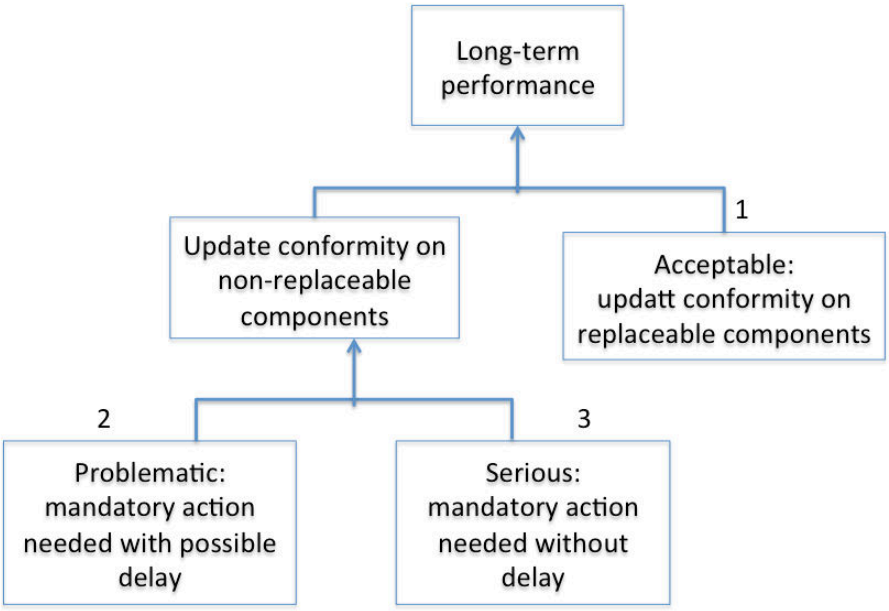

Figure A.4 “Scoring” of criticality for criterion “Level of Long-term performance”

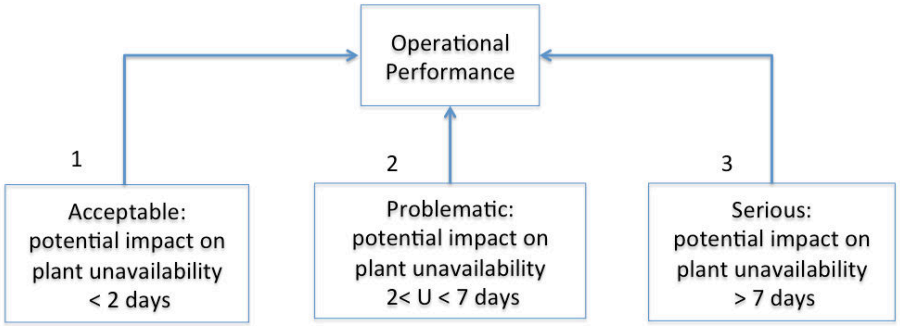

Figure A.5 "Scoring" of criticality for criterion "Level of Operational performance"

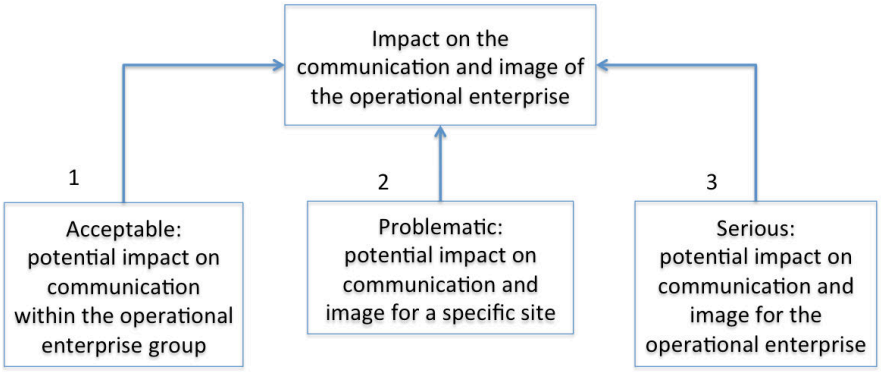

Figure A.6 "Scoring" of criticality for criterion "Level of Impact on the Communication and Image of the Operational Enterprise"

APPENDIX B. Mathematical details about the algorithm of disaggregation of a MR-Sort classification model

We consider the case involving k categories that are, thus, separated by *(k-1)* frontier denoted $b = \{b^1, b^2, ..., b^h, ..., b^{k-1}\}$, where $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}, h = 1, 2, ..., k$, n is the number of criteria that are taken into account. Let $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N_{TR}\}$ be the training set, where N_{TR} is the number of alternatives and, $(A^1, A^2, ..., A^k)$ be the partition of the training set, ordered from the "best" to the "worst" classes.

For each alternative $x_p \in D_{TR}$, in category A^h of the learning set D_{TR} (for $h = 2, 3, ..., k-1$), let us define *2n* binary variables δ_{ip}^h and δ_{ip}^{h-1}, for $p = 1, 2, ..., N_{TR}$, such that δ_{ip}^l equals to 1 iff $g_i(x_p) \geq b_i^l$ for $l = h-1, h$ and $\delta_{ip}^h = 0 \Leftrightarrow g_i(x_p) < b_i^h$. We introduce *2n* continuous variables c_{ip}^l $(l = h-1, h)$ constrained to be equal to ω_i if $\delta_{ip}^l = 1$ and to *0* otherwise.

We consider an objective function that describes the robustness of the assignment. We introduce two more continuous variables, y_p and z_p, for each $x_p \in D_{TR}$ and α. In maximizing α, we maximize the value of the minimal slack in the constraints.

We resume all the constraints in the following mathematical program:

$$
\begin{aligned}
&max\ \alpha,\\
&\alpha \le y_p, \alpha \le z_p, \forall x_p \in D_{TR},\\
&\sum_{i,p} c_{ip}^l + y_p + \epsilon = \lambda, \forall x_p \in A^{l-1},\\
&\sum_{i,p} c_{ip}^l = \lambda + z_p, \forall x_p \in A^l,\\
&c_{ip}^l \le \omega_i, \forall x_p \in D_{TR},\\
&c_{ip}^l \le \delta_{ip}^l, \forall x_p \in D_{TR},\\
&c_{ip}^l \ge \delta_{ip}^l - 1 + \omega_i, \forall x_p \in D_{TR},\\
&M\delta_{ip}^l + \epsilon \ge g_i(x_p) - b_i^l, \forall x_p \in D_{TR},\\
&M(\delta_{ip} - 1) \le g_i(x_p) - b_i^l, \forall x_p \in D_{TR},\\
&\sum_{i,p} \omega_i = 1, \lambda \in [0.5, 1],\\
&\omega_i \in [0, 1],\\
&c_{ip}^l \in [0, 1], \delta_{ip}^l \in \{0, 1\}, \forall x_p \in D_{TR}\\
&y_p, z_p \in \mathbb{R}, \forall x_p \in D_{TR},\\
&i \in \{1, 2, ..., n\}, p \in \{1, 2, ..., N_{TR}, \},\\
&\alpha \in \mathbb{R}.
\end{aligned}
$$

M is an arbitrary large positive value and ε an arbitrary small positive quantity.

The case in which x_p belongs to one of the extreme categories (A^1 and A^k) is simple. It requires the introduction of only *n* binary variables and *n* continuous variables. In fact, if x_p belongs to A^1 we just have to express that the subset of criteria on which x_p is at least as good as b_1 has sufficient weight. In a dual way, when x_p lies in A^k, the worst category, we have to express that it is at least as good as b_k on a subset of criteria that has not sufficient weight.

In case it is infeasible to generate a compatible model based on the given data, an inconsistency study of the data set is demanded (see Section 5). The formulation above is, then, modified accordingly as follows:

B.1 The constraints deletion algorithm

For each $x_p \in D_{TR}$, as presented in Section 5.1, we introduce a binary variable γ_p, which is equal to "1" if alternative x_p is correctly assigned by the MR-sort model, and equal to "0" otherwise. To ensure that the γ_p variables are correctly defined, the following constraints (2) of the basic disaggregation algorithm listed above

$$\sum_{i,p \in N} c_{ip}^{l} + y_p + \epsilon = \lambda, \forall x_p \in A^{l-1},$$
$$\sum_{i,p \in N} c_{ip}^{l} = \lambda + z_p, \forall \in A^{l}; \qquad (2)$$

are replaced by

$$\sum_{i \in N} c_{ip} < \lambda + M(1-\gamma_p), \forall x_p \in D_{TR}^{l-1},$$
$$\sum_{i \in N} c_{ip} \geq \lambda - M(1-\gamma_p), \forall x_p \in D_{TR}^{l}; \qquad (3)$$

Correspondingly, the objective function becomes

$$\alpha = \sum_{x_p \in D_{TR}} \gamma_p. \qquad (4)$$

By so doing, we obtain a MIP that yields a subset $D_{TR}^{*} \subseteq D_{TR}$ of maximal cardinality that can be represented by an MR-sort model.

B.2 The constraints relaxation algorithm

As presented in Section 3, we consider the case involving *k* categories that are, thus, separated by *(k-1)* frontier denoted $b = \{b^1, b^2, ..., b^h, ..., b^{k-1}\}$, where $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}, h = 1, 2, ..., k$, *n* is the number of criteria that are taken into account. For each alternative $x_p \in D_{TR}$, its predefined category is given by Γ_p. For the ones that are assigned to the extreme categories, i.e., the best category *1* and the worst category *k-1*, only one constraint can be obtained, which shows that with respect to the best (resp., worst) set of lower profiles $b^1 = \{b_1^1, b_2^1, ..., b_n^1\}$ (resp., $b^{k-1} = \{b_1^{k-1}, b_2^{k-1}, ..., b_n^{k-1}\}$), they are even better (resp., worse). For the rest of the alternatives that are assigned to the intermediate categories, two constraints are gathered, e.g., alternative $\{x_p | \Gamma_p = t, 1 < t < k-1, t \in \mathbb{N}\}$ should be better with respect to the lower profile set $b^t = \{b_1^t, b_2^t, ..., b_n^t\}$, and worse with respect to the lower profile set $b^{t-1} = \{b_1^{t-1}, b_2^{t-1}, ..., b_n^{t-1}\}$.

As mentioned in Section 5.2, we introduce binary variables γ_p^+ (regarding to the better category's lower profile set, except for $\Gamma_p = 1$) and γ_p^- (referring to the worse category's lower profile set, except for $\Gamma_p = k$), which are equal to "1" if alternative x_p fulfills the constraints comparing with the corresponding lower profile sets, "0" otherwise. We modify the algorithm in the previous subsection (5.1) in the following way:

$$\begin{aligned} &\sum_{i\in\mathbb{N}} c_{ip} < \lambda + (1-\gamma_p^+), \forall x_p \in D_{TR}^{l-1}, \\ &\sum_{i\in\mathbb{N}} c_{ip} \geq \lambda - (1-\gamma_p^-), \forall x_p \in D_{TR}^{l}; \end{aligned} \qquad (5)$$

Correspondingly, the objective function (4) is replaced by the new objective

$$\alpha = \sum_{x_p \in D_{TR}} \gamma_p^+ + \sum_{x_p \in D_{TR}} \gamma_p^- \qquad (6)$$

We, thus, obtain a MIP that yields a subset $D_{TR}^* \subseteq D_{TR}$ of maximal cardinality that can generate an MR-sort model. In addition, for each of the alternatives that are not accepted into the subset D_{TR}^*, the corresponding inconsistent constraints are also targeted: for example, if for one alternative $\{x_p | \Gamma_p = t, 1 < t < k-1, t \in \mathbb{N}\}$ we obtain $\gamma_p^+ = 0$ (resp., $\gamma_p^- = 0$), then this alternative should be classified in a better (resp., worse) category with respect to the lower profile set of category *t-1* (resp., *t*); in other words, its original assignment is underestimated (resp., overestimated). The same criterion is applied to the alternatives that are originally assigned to the best or worst category.

REFERENCES

1 Wang Q, Poh K.L. A survey of integrated decision analysis in energy and environmental modeling. Energy, 2014; 77, pp. 691-702.

2 Aven T. Foundations of Risk Analysis. Germany, Berlin: Wiley, N.J, 2003.

3 Aven T. Some reflections on uncertainty analysis and management. Reliability Engineering and System Safety, 2010; 95, pp. 195-201.

4 Kröger W, Zio E. Vulnerable Systems. UK, London: Springer, 2001.

5 Huang J.P., Poh K.L. and Ang B.W. Decision analysis in energy and environmental modeling. Energy,

1995; 20, pp. 843-855.

6 Leroy A, Mousseau V, Pirlot M. Learning the parameters of a multiple criteria sorting method, The Second International Conference on Algorithmic Decision Theory, Algorithmic Decision Theory, R.I. Brafman, F. Roberts, and A. Tsoukiàs (Eds.): ADT 2011, LNAI 6992, pp. 219–233, Germany, Berlin: Springer, 2011

7 Wang T-R, Mousseau V, Zio E. A hierarchical decision making framework for vulnerability analysis. pp. 1–8. Proceedings of ESREL2013, Amsterdam, The Netherlands, 2013.

8 Roy B. The outranking approach and the foundations of ELECTRE methods. Theory and Decision 31, 1991, pp. 49-73.

9 Mousseau V., Slowinski R. Inferring an ELECTRE TRI Model from Assignment Examples. Journal of Global Optimization, vol. 12, 1998, pp. 157-174.

10 Aven T, Flage R. Use of decision criteria based on expected values to support decision-making in a production assurance and safety setting. Reliability Engineering and System Safety, 2009; 94, pp. 1491-1498.

11 Milazzo MF, Aven T. An extended risk assessment approach for chemical plants applied to a study related to pipe ruptures. Reliability Engineering and System Safety, 2012; 99, pp. 183-192.

12 Rocco C, Zio E. Bootstrap-based techniques for computing confidence intervals in Monte Carlo system reliability evaluation. pp. 303–307. Proceedings of the Annual Reliability and Maintainability Symposium, 2005. IEEE

13 Baraldi, P., Razavi-Far, R., Zio, E., 2010. A Method for Estimating the Confidence in the Identification of Nuclear Transients by a Bagged Ensemble of FCM Classifiers. Seventh American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies NPIC&HMIT 2010, Las Vegas, Nevada, November 7-11, 2010, on CD-ROM, American Nuclear Society, LaGrange Park, IL (2010).

14 Doumpos M., Zopounidis C., Multricriteria Decision Aid Classification Methods, Kluwer Academic

Publishers, Netherlands. 2002, ISBN 1- 4020-0805-8.

15 NWRA, N. W. R. A. Risk assessment methods for water infrastructure systems, 2012. Rhode Island Water Resources Center, University of Rhode Island, Kingston, RI.

16 Mousseau V., Dias C.L., Figueira J. Dealing with inconsistent judgments in multiple criteria sorting models. 4OR: A Quarterly Journal of Operations Research, 2005; 4, pp. 145-158.

17 Baraldi P, Razavi-Fra R, Zio E. Bagged ensemble of fuzzy C means classifiers for Nuclear Transient Identification. Annals of Nuclear Energy, Elsevier Masson, 2011, 38, pp. 1161-1171.

18 Wilson R, Martinez TR. Combining cross-validation and confidence to measure fitness. Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), 1999; pp. 1409-1416. Washington D.C.IEEE

19 Gutierrez-Osuna, R. Pattern analysis for machine olfaction: A review. IEEE SENSORS JOURNAL, 2002, 10.1109/JSEN.2002.800688

20 Efron B, Thibshirani RJ. An introduction to the bootstrap. Monographs on statistics and applied probability 57, 1993. Chapman and Hall, New York.

21 Zio E, A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. IEEE Transactions on Nuclear Science, 53(3), 2006; pp. 1460-1470.

22 Cadini F, Zio E, Kopustinskas V, Urbonas R. An empirical model based bootstrapped neural networks for computing the maximum fuel cladding temperature in a RBMK-1500 nuclear reactor accident. Nuclear Engineering and Design, 238, 2008; pp. 2165-2172.

Paper (iv)

Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent intentional acts: a sensitivity-based decision-making approach

T. R. Wang, N. Pedroni, and E. Zio.

Reliability Engineering & System Safety, 2015, accepted.

Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent acts: a sensitivity-based decision-making approach

Tai-Ran WANG[a], Nicola PEDRONI[a], Enrico ZIO[a,b]

a Chair on Systems Science and the Energy challenge, Fondation EDF, Ecole Centrale Paris and Supelec, Grande Voie des Vignes, F92-295, Chatenay Malabry Cedex

b Politecnico di Milano, Energy Department, Nuclear Section, c/o Cesnef, via Ponzio 33/A , 20133, Milan, Italy, Fax: 39-02-2399.6309, Phone: 39-02-2399.6340, enrico.zio@polimi.it

ABSTRACT

A classification model based on the Majority Rule Sorting method has been previously proposed by the authors to evaluate the vulnerability of safety-critical systems (e.g., nuclear power plants) with respect to malevolent intentional acts.

In this paper, we consider a classification model previously proposed by the authors based on the Majority Rule Sorting method to evaluate the vulnerability of safety-critical systems (e.g., nuclear power plants) with respect to malevolent intentional acts. The model is here used as the basis for solving an inverse classification problem aimed at determining a set of protective actions to reduce the level of vulnerability of the safety-critical system under consideration.

To guide the choice of the set of protective actions, sensitivity indicators are originally introduced as measures of the variation in the vulnerability class that a safety-critical system is expected to undergo after the application of a given set of protective actions. These indicators form the basis of an algorithm to rank different combinations of actions according to their effectiveness in reducing the safety-critical systems vulnerability. Results obtained using these indicators are presented with regard to the application of: (i) one identified action

at a time, (ii) all identified actions at the same time or (iii) a random combination of identified actions. The results are presented with reference to a fictitious example considering nuclear power plants as the safety-critical systems object of the analysis.

KEYWORDS: safety-critical system, malevolent intentional attacks, vulnerability analysis, protective actions, Majority Rule Sorting (MR-Sort), classification model, inverse classification problem, sensitivity indicator

Notations

$crit^j$	subcriterion j
x_i	main criterion i
NPP_i	Nuclear power plant i
C_i	vulnerability category i
act^k	protective action k
$coeff^{kj}$	weight of the influence of action k on attribute j
$crit'^j$	after action subcriterion j
B	limited budget
$N\uparrow$	number of NPPs that are improved after the action(s)
$\frac{N\uparrow}{N'}$	estimate of the percentage of new NPPs that can be expected to be improved
$N\downarrow$	number of NPPs that are expected to be deteriorated after the action(s)
$\frac{N\downarrow}{N'}$	estimate of the percentage of new NPPs that can be expected to be deteriorated
$\Delta N = \frac{N\uparrow}{N'} - \frac{N\downarrow}{N'}$	expected "net" amount of ameliorated NPPs
$\Delta M\uparrow$	total variation of category underwent by the ameliorated

	NPPs
$\frac{\Delta M\uparrow}{N\prime}$	variation in vulnerability category that a new ameliorated plant is expected to undergo
$\Delta M\downarrow$	total variation of category underwent by the deteriorated NPPs
$\frac{\Delta M\downarrow}{N\prime}$	variation in vulnerability category that a new deteriorated plant is expected to undergo
$\overline{\Delta M} = \frac{\Delta M\uparrow}{N\prime} - \frac{\Delta M\downarrow}{N\prime}$	"net" variation in vulnerability category that a newly analyzed NPP is expected to undergo.
$\Delta S\uparrow$	ratio between the sums of the variations of vulnerability category underwent by the ameliorated NPPs and the sum of the corresponding maximum possible category variations
l_i^j	level of action j applied on system i

1. INTRODUCTION

The vulnerability of safety-critical systems and infrastructures (e.g., nuclear power plants) is of great concern, given the multiple and diverse hazards that they are exposed to (e.g., intentional, random, natural etc.) [1] and the potential large-scale consequences. This justifies the increased attention for analyses aimed at (i) the systematic identification of the sources of system vulnerability, (ii) the qualitative and quantitative assessment of system vulnerability [2][3] and (iii) the definition of effective actions of vulnerability reduction.

In a previous work [6], we have proposed an empirical classification framework to tackle the issue (ii) of assessing vulnerability to malevolent intentional acts. Specifically, we have adopted a classification model based on the Majority Rule Sorting (MR-Sort) method [7] to

assign an alternative (i.e., a safety-critical system) to a given (vulnerability) class (or category). The MR-Sort classification model contains a group of (adjustable) parameters that are calibrated by means of a set of empirical classification examples (also called training set), i.e., a set of alternatives with the corresponding pre-assigned vulnerability classes [6][7]. For further details on this method, the interested reader can refer to the Appendix A at the end of the paper. It is worth mentioning that other majority rule voting methods are widely used in technical decision making problems for vulnerability analysis of systems, see, e.g., [21].

In this paper, we are still only concerned with intentional hazards (i.e., those related to malevolent acts) and address issue (iii) above (i.e., the definition of the actions to undertake for reducing the level of system vulnerability). This issue is difficult to be resolved by traditional risk assessment methods [1][4][5]. On the contrary, the base model developed in Ref. [6] can be extended to address the problem relates to the problem of optimal risk reduction, e.g. by optimization of protective measures [29][30][31]. In other words, an *inverse classification* problem [8][9][10] of determining a set of protective actions that can effectively reduce the level of vulnerability of a safety-critical system [11], taking into account a specified set of constraints (e.g., budget limits) [8].

The present analysis can be considered part of an encompassing business process of safety management (see, e.g., [22]), where we seek for the best compromise among risks, costs and benefits in allocating investments in safety-critical systems in the presence of uncertainties [28]. Correspondingly, the presented algorithms can be considered part of an encompassing business process of safety management [22]. Mathematically speaking, the aim is to identify how to modify some features of the input patterns (i.e., the attributes of the safety-critical system under analysis) such that the resulting class is changed as desired (i.e., the vulnerability category is reduced to a desired level). To achieve this objective, novel sensitivity indicators [12] are introduced for quantifying the variation in the vulnerability class of a safety-critical system resulting from the application of a given set of protective actions [13]. Using these indicators as the basis for a ranking algorithm, changes in system vulnerability can be achieved considering: (i) one identified action at a time, (ii) all identified

actions at the same time or (iii) a random combination of identified actions. The proposed indicators also allow different combinations of actions to be ranked and their effectiveness in reducing the vulnerability under specified budget constraints can be evaluated on a new (test) set of (unknown) safety-critical systems (i.e., systems not used before to calibrate/train the classification model). In this context, it is known that existing risk assessment methodologies may fail to account for unknown and emergent risks that are typical of large-scale infrastructure investment allocation problems. On the other hand, in modern portfolio theory, it is well known that a diversified portfolio can be very effective to reduce non-systematic risks. The approach of diversification is equally important in choosing robust portfolios of infrastructure projects that may be subject to emergent and unknown risks [27]. The proposed methodology is expected to contribute also in this direction of optimal classification of options/investments and combinations of the same.

The remainder of the paper is structured as follows. Section 2 recalls the modeling framework for the analysis of vulnerability to intentional hazards. With reference to that, Section 3 introduces the problem of inverse classification. Section 4 describes the sensitivity analysis indicators introduced to tackle the inverse classification problem of Section 3. Section 5 illustrates their use for the identification of protective actions. In Section 6, a case study is proposed to show the application of the method. Finally, Section 7 gives the discussion and conclusions of this research.

2 THE CLASSIFICATION MODEL FOR THE ASSESSMENT OF VULNERABILITY TO INTENTIONAL HAZARDS

We limit the vulnerability analysis of a system to the evaluation of the susceptibility to intentional hazards and adopt the three-layers hierarchical model developed in [6] (Figure 1). The susceptibility to intentional hazards (level 1 in Figure 1) is characterized in terms of attractiveness and accessibility (level 2 in Figure 1). These attributes are hierarchically broken down into factors which influence them, including resilience interpreted as pre-attack

protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness). The disaggregation is made in 6 criteria (level 3 in Figure 1): physical characteristics (x1), social criticality (x2), possibility of cascading failures (x3), recovery means (x4), human preparedness (x5) and level of protection (x6). These six criteria are further decomposed into a layer of m=16 basic subcriteria $\{crit^j, j = 1,2, \dots, m = 16\}$(level 4 in Figure 1), for which data and information are collected in terms of quantitative values or linguistic terms depending on the nature of the subcriterion. The descriptive terms and/ or values of the fourth layer subcriteria are, then, scaled to numerical categories. The criteria included in the layers are defined and assigned "preference directions" for treatment in the decision-making process. The preference direction for a given criterion (e.g., a physical characteristic or parameter of the system) indicates the state towards which it is desirable to "move the parameter" in order to reduce system susceptibility: in other words, the preference direction is assigned from the point of view of a "defender" who is concerned with protecting the system from an attack [16]. Finally, to get the value of the six third-layer criteria $\{x_{i,} i = 1,2,..,6\}$, (i) we assign weights to each subcriterion to indicate its importance and (ii) we apply a simple weighted sum to the categorical values of the constituent subcriteria $\{crit^j, j = 1,2, \dots, m = 16\}$. These *m*=16 criteria $\{crit^j, j = 1,2, \dots, m = 16\}$ are evaluated to assess the vulnerability of a given safety-critical system of interest (e.g., a nuclear power plant – NPP).

For the purposes of the present analysis, *M* = 4 levels (or categories) of system vulnerability $\{Class = C, C = 1,2,3,4\}$ are considered: 1 = satisfactory, 2 = acceptable, 3 = problematic and 4 = serious.

Then, the assessment of vulnerability corresponds to a classification problem: given the definition of the characteristics of a critical system in terms of the sixteen criteria above, assign the vulnerability category (or class) to which the system belongs.

The classification model is based on the Majority Rule Sorting (MR-Sort) method [7][14][15]; the model contains a group of (adjustable) parameters that have to be calibrated by means of a set of empirical classification examples (the training set), i.e., a set of

alternatives with the corresponding pre-assigned vulnerability classes. Further details about the classification model are not reported here for brevity: the interested reader is referred to [16].

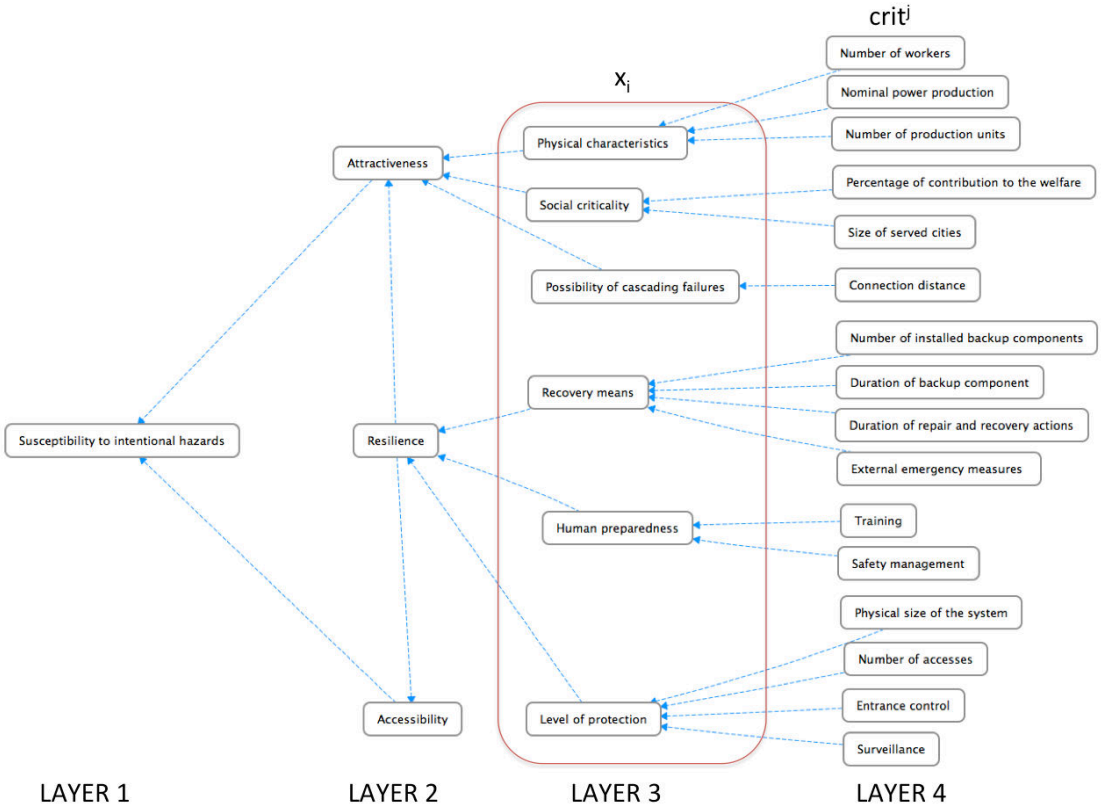

Figure 1. Hierarchical model for susceptibility to intentional hazards [16]

3. INVERSE CLASSIFICATION PROBLEM FOR PROTECTIVE ACTIONS IDENTIFICATION

We define an inverse classification problem aimed at finding a combination of actions reducing the vulnerability of a (group of) safety-critical system(s) eventually under budget limitations.

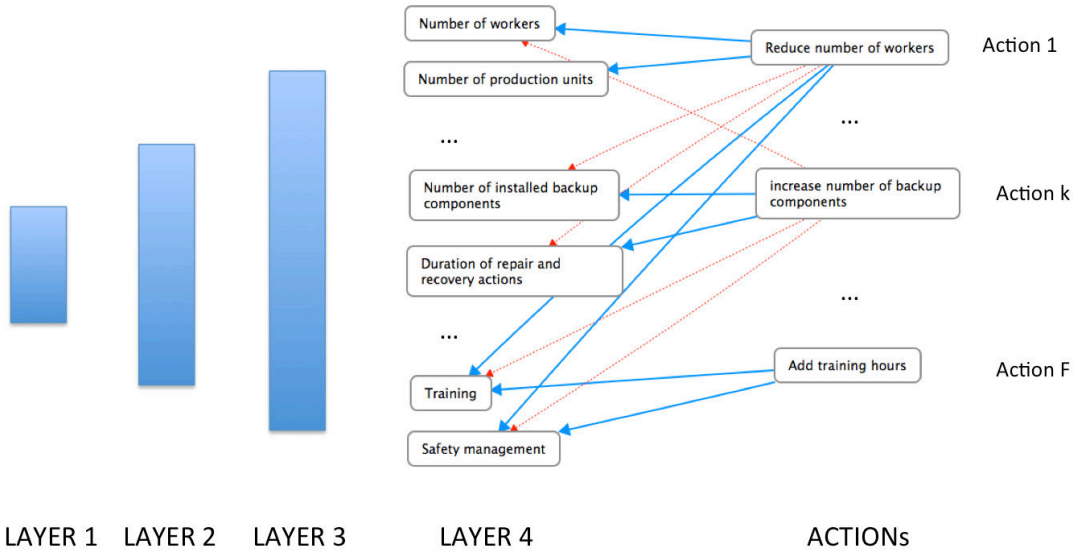

Figure 2. Schema of direct actions for basic criteria

To illustrate the methodology, we consider a set of N NPPs ($NPP_i, i \in \{1,2,\dots,N\}$) characterized by $m = 16$ basic features ($crit^j, j \in \{1,2,\dots,m\}$). On the basis of these m=16 features, the NPPs are assigned to M=4 *pre-defined* categories ($C_i \in \{1,2,\dots,M\}, i \in \{1,2,\dots,N\}$), where $C_i = 1$ represents the best situation, i.e., lowest vulnerability. Let $act = \{act^1, act^2, \dots, act^F\}$ denote the available set of actions, each of which can influence one or more basic criteria $crit^j, j \in \{1,2,\dots,m\}$ (Figure 2) with different intensity, as measured by a set of coefficients $\{coeff^{kj}, k \in \{1,2,\dots,F\}, j \in \{1,2,\dots,m\}\}$. In other words, $coeff^{kj}$ is the "weight" of the influence of action k on attribute j (the higher the absolute value of $coeff^{kj}$, the stronger the effect of action k on attribute j). Notice that a positive (resp. negative) coefficient $coeff^{kj}$ means that action k has an ameliorative, positive (resp. deteriorative, negative) effect on attribute j, that is, it changes the corresponding value towards (resp. away from) the "preference direction" of attribute j; on the contrary, if $coeff^{kj}$ is equal to zero, then criterion j is not influenced by action k. "Negative" relationships objectively exist. Actually, taking one action to improve the performance of one specific criterion may lead to a "negative" change in some of the others. For example, increasing the number of backup components on site may lead to an increased number of workers to operate and maintain them, which may increase the possibility of a larger number

of injuries of the people exposed to an attack. If the analyst who builds the inverse classification model were not able to identify and quantify these "negative" connections (i.e., the coefficients $coeff^{kj}$), then the (positive) effect of a given combination of actions on a system could be overestimated, with serious drawbacks on the process of resources allocation for system protection.

Significant efforts have been made to assign numerical values to the impacts of actions, in order to represent the problem as realistically as possible. However in a non-fictitious situation, the task is expected to be more complex. Actually, the relations between the actions and the criteria taking into account the dependencies of different attributes and systems are always difficult to identify: in such cases, resorting to the judgment of real experts and possibly to real historical data will be mandatory.

The implementation of one or more actions modifies the attribute values $crit^j, j \in \{1,2,\dots,m\}$ and as a result, the vulnerability of the system (i.e., the assignment by the classification model) may change. In this paper, we assume that the total effect of the available set of actions $act = \{act^1, act^2, \dots, act^F\}$ on criterion j is obtained by a linear superposition of the effects of each action act^i:

$$crit'^j = crit^j + \sum_{k=1}^{F} coeff^{kj} * act^k;$$

$$k \in \{1,2,\dots,F\}, j \in \{1,2,\dots,m\}. (1)$$

where $crit'^j$ is the value of attribute j after the identified set of available actions has been implemented.

Also, let $Cost(NPP_i, act'), act' \subseteq act$ denote the cost of the combination of actions act' applied to NPP_i. The inverse classification problem can then be formulated as follows: identify the set of actions $act'_i (\subseteq act), i = 1,2,\dots,N$ that improve the vulnerability of the system to a demanded vulnerability category C_i^λ while minimizing the cost, i.e.,

$$\min\left(\sum_{i=1}^{N} Cost(NPP_i, act'_i)\right), act'_i \subseteq act;$$

$$s.t. classify(NPP_i, act'_i) = C_i^\lambda;$$

$$i \in \{1,2,\dots,N\}(2)$$

Alternatively, if it is known that the budget B_i is limited for each plant NPP_i, the formulation becomes: improve the systems to the best possible vulnerability category $C_i^{\lambda\prime}, C_i^{\lambda\prime} \in \{1,2,\dots,M\}, i \in \{1,2,\dots,N\}$, while keeping the cost below the available budget B:

$$\max\big(classify(NPP_i, act'_i)\big);$$

$$s.t.\, Cost(NPP_i, act'_i) \leq B_i, act'_i \subseteq act;$$

$$i \in \{1,2,\dots,N\}. (3)$$

To address the inverse classification problem, we adopt a pragmatic approach based on sensitivity analysis [17][18][19], introducing indicators that quantify the variation in the vulnerability class that a safety-critical system is expected to undergo upon implementation of a given set of actions.

4 SENSITIVITY INDICATORS FOR DRIVING THE INVERSE CLASSIFICATION PROBLEM

We consider the group of N' vulnerability-class labeled known (available) safety-critical systems (NPPs) used to train the MR-Sort classification model and study the sensitivity of their categories of vulnerability to the implementation of the available protective actions. We denote the original categories of these NPPs as $C_i, C_i \in \{1,2,\dots,M\}, i \in \{1,2,\dots,N'\}$ and the new categories resulting from the application of a set of protective actions as $C_i^{\lambda}, C_i^{\lambda} \in \{1,2,\dots,M\}, i \in \{1,2,\dots,N'\}$.

Let $N\uparrow$ be the number of NPPs that are improved after the action(s):

$$N\uparrow = \sum_{i=1}^{N'} A_i, i \in \mathbb{N};$$

$$A_i = 1, if\ C_i > C_i^{\lambda};$$

$$A_i = 0, if\ C_i < C_i^{\lambda}. (4)$$

Then, $\frac{N\uparrow}{N\prime}$ can be interpreted as an estimate of the percentage of new (i.e., different from the

ones of the training set) NPPs that can be expected to be improved after such action(s) is (are) implemented on them.

Dually, $N\downarrow$, is the number of NPPs that are expected to be deteriorated after the action(s):

$$N\downarrow = \sum_{i=1}^{N'} A_i, i \in \mathbb{N};$$

$$A_i = 1, if\ C_i < C_i^{\lambda};$$

$$A_i = 0, if\ C_i > C_i^{\lambda}. (5)$$

Notice that a "deterioration" (i.e., an increase in the vulnerability category) is possible because some of the actions may have positive effects on some subcriteria but negative effects on some others (see Section 3). Then, $\frac{N\uparrow}{N\prime}$ can be interpreted as an estimate of the percentage of new NPPs (i.e., different from the ones of the training set) that can be expected to be deteriorated after such action(s) is (are) implemented on them.

We consider the quantity $\Delta N = \frac{N\uparrow}{N\prime} - \frac{N\downarrow}{N\prime}$ to combine the effects of both positive and negative influences of the actions in the expected "net" amount of ameliorated NPPs.

Considering that the evaluation framework is based on *M*=4 categories, it seems reasonable to consider not only the number of NPPs that are ameliorated or deteriorated, but also the amount of variation in category of vulnerability of each of them. To this aim, we introduce the following indicators to combine the amount of variation in vulnerability with the number of NPPs whose vulnerability category has changed after the actions.

In particular, $\Delta M\uparrow$ is defined as the total variation of category underwent by the *ameliorated* NPPs:

$$\Delta M\uparrow = \sum_{i=1}^{N'} M_i * A_i, i \in \mathbb{N};$$

$$M_i = C_i - C_i^{\lambda};$$

$$A_i = 1, if\ C_i > C_i^{\lambda};$$

$$A_i = 0, if\ C_i < C_i^{\lambda}. (6)$$

Thus, $\frac{\Delta M\uparrow}{N\prime}$ can be interpreted as the variation in vulnerability category that a new *ameliorated* plant is expected to undergo when the chosen combination of actions is applied.

Dually, $\Delta M\downarrow$ is defined as:

$$\Delta M\downarrow = \sum_{i=1}^{N'} M_i * A_i, i \in \mathbb{N};$$

$$M_i = C_i - C_i^{\lambda};$$

$$A_i = 1, if\ C_i < C_i^{\lambda};$$

$$A_i = 0, if\ C_i > C_i^{\lambda}. (7)$$

Thus, $\frac{\Delta M\downarrow}{N\prime}$ can be seen as the variation in vulnerability category that a new *deteriorated* plant is expected to undergo when the chosen combination of actions is applied.

Finally, $\overline{\Delta M} = \frac{\Delta M\uparrow}{N\prime} - \frac{\Delta M\downarrow}{N\prime}$ combines the effects of both positive and negative influences of the actions and it can be seen as the "net" variation in vulnerability category that a newly analyzed NPP is expected to undergo after the application of the given set of actions.

The net expected variation in vulnerability category $\overline{\Delta M}$ quantifies the influence of the actions upon the NPPs. However, this measure does not take into account the original category assignment of the NPPs: for example, in practice there is a difference between taking a NPP from category 4 to 3 and taking it from 2 to 1, even if the category variation is 1 in both cases. To consider this, we introduce the indicator $\Delta S\uparrow$, defined as the ratio between the sums of the variations of vulnerability category underwent by the ameliorated NPPs and the sum of the corresponding maximum possible category variations (i.e., the sum of the category variations that the NPPs would undergo if they were ameliorated to the best possible vulnerability category):

$$\Delta S\uparrow = \frac{\Delta M\uparrow}{E};$$

$$E = \sum_{i=1}^{N\prime} \left(C_i - C^{best}\right) * A_i, i \in \mathbb{N};$$

$$A_i = 1, if\ C_i > C_i^{\lambda};$$

$$A_i = 0, if\ C_i < C_i^{\lambda}. (8)$$

The indicator $\Delta S \uparrow$ quantifies the influence of the actions on NPPs, relative to their original categories: the lower $\Delta S \uparrow$ is, the higher the influence of the chosen set of actions is on the NPPs originally of a relatively low category.

Based on the above indicators, an algorithm is proposed to rank different combinations of actions according to their effectiveness in reducing the vulnerability of safety-critical systems. The actions with positive influences are obviously preferred. On the contrary, concerning the ones with negative influences, the rationality of being chosen as ameliorative actions should be reconsidered. The analyst may replace/modify/delete them from the original considered action set. The algorithm proceeds as follows:

(1) Rank the (combinations of) actions according to the value of $\overline{\Delta M}$ (the higher the value of $\overline{\Delta M}$, the more effective the combination of actions in reducing vulnerability):

- combinations of actions that have a negative value of $\overline{\Delta M}$ ($\overline{\Delta M} < 0$) are expected to increase the vulnerability of a NPP: this is due to the fact that some actions may have a deteriorated effect on some of the subcriteria that more than counter balances the positive effects on their subcriteria. The identification of the combination of actions with $\overline{\Delta M} < 0$allows the analyst to (i) study the mechanisms of influence of the actions on the basic subcriteria (Layer 4 in Figures 1 and 2) and (ii) if possible, eliminate the "negative connections", i.e., the negative dependencies between some actions and some criteria (e.g., by identifying alternative actions for dealing with these "critical" subcriteria);
- the actions that have a positive value of $\overline{\Delta M}$ ($\overline{\Delta M} > 0$) are expected to reduce the vulnerability and are assigned higher rankings (the higher $\overline{\Delta M}$, the higher the ranking);

(2) If several combinations of actions have the same value of $\overline{\Delta M}$, then consider the other indicators (i.e., $\frac{N\uparrow}{N\prime}$ and $\frac{\Delta M\uparrow}{N\prime}$): depending on the judgment of the DMs, higher importance may

be given to those actions that produce a larger expected number of improved NPPs ($\frac{N\uparrow}{N\prime}$) or to those that generate a higher "expected class improvement" ($\frac{\Delta M\uparrow}{N\prime}$).

(3) If some combinations still have the same ranking, analyze indicator $\Delta S\uparrow$ to check which actions have stronger impact on the NPPs of low categories.

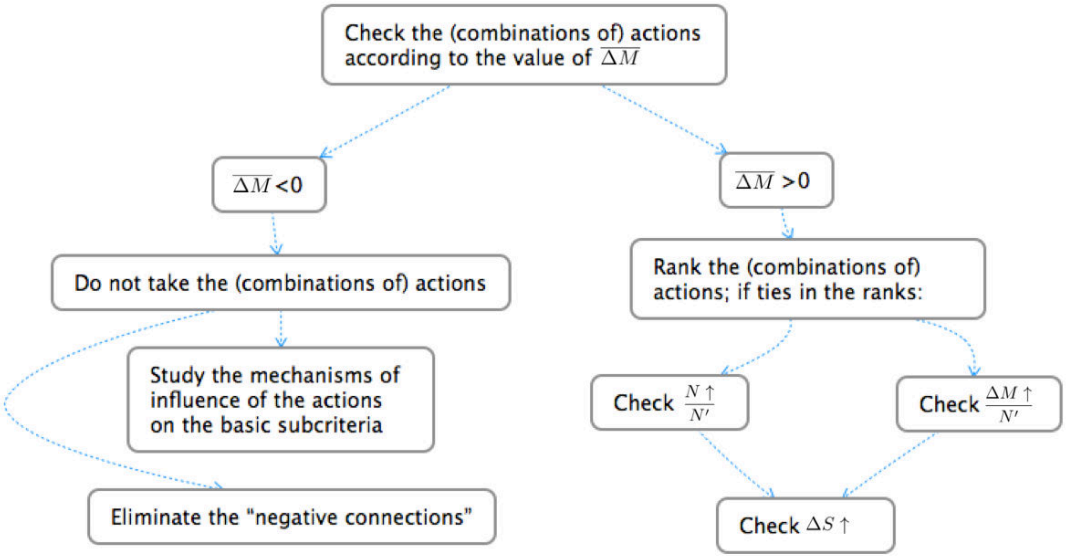

Figure 3. Schema of decision logic for selecting an action

5 CASE STUDY

The sensitivity analysis proposed in Section 4 is applied on a case study concerning the vulnerability analysis of NPPs [6]. We refer to the n=6 main criteria $i = 1,2,\dots,n = 6$ of the hierarchical modeling presented in [6] and recalled in Section 2: physical characteristics (x_1), social criticality (x_2), possibility of cascading failures (x_3), recovery means (x_4), human preparedness (x_5) and level of protection (x_6); these criteria are numbers scaled in the range [0,1]. Then, the main criteria are successively broken into a layer of m=16 basic criteria (Figure 2). Finally, M=4 vulnerability categories $Class = C, C = 1,2,3,4$ are defined as: 1= satisfactory, 2= acceptable, 3= problematic and 4 = serious (Section 2).

As shown in Figure 2 and anticipated in Section 3, we define F=13 direct actions ($act = \{act^1, act^2, \dots, act^F\}$), each acting on one or more subcriteria (Table 1). All the actions have multiple influences on different criteria, with possibly positive or negative effects: for

example, the action “reduce the number of workers” has an obvious direct influence on the subcriterion “Number of workers”, but may also imply, e.g., (i) reducing the number of production units, the number of accesses to the plant, the number of installed backup components and external emergency measures; (ii) increasing the duration of repair and recovery actions; (iii) enhancing the training; (iv) facilitating the safety management and entrancing control and surveillance. The strengths of the influences of the actions on the different criteria are quantified by the different weights/coefficients reported in Table 1.

Also, for each action we consider different levels of implementation l_i^j ($l_i^j, l_i^j \in \{0,1,2,3\}, i \in \{1,2,\dots,N\}, j \in \{1,2,\dots,F\}$), representing to what extent/ how far/ in which amount action *j* is applied on system *i* (notice that $l_i^j = 0$ means that action *j* is not applied to system *i*) (Table 1).

Finally, for simplicity we assume that the cost related to the application of a given action is equal to the level l_i^j of the action: for example, referring to Table 1, if we choose to reduce the number of workers by 20%, the related cost is 1 in arbitrary units (since the action corresponds to level l_i^j=1); on the contrary, if we reduce the number of workers by 30%, the cost is 3 (since the action corresponds to level l_i^j=3). The idea is that the cost of an action increases (resp. decreases) with its “level” of “strength” of implementation. Notice that however, the levels assigned to the actions are not always strictly “mathematically” proportional to the change of value they produce in the corresponding criteria. In fact, for different actions, the three levels of “effects” on the corresponding directly influenced criteria may be of different notice. Sometimes they may be represented by a quantitative discrete number (e.g., for action “reduce number of production units” we have -1 production unit for level 1, -2 production units for level 2, -3 production units for level 3); sometimes they may be a percentage (as for the number of workers mentioned above). In addition, the costs of an action and the corresponding change in a criterion value are not strictly proportional either (e.g., the cost of training enhance may be the same for 50 and for 80 people, but different for 100). In this view, choosing the cost of an action equal to the level l_i^j of implementation of

the action is a (maybe rough) compromise between simplicity and pragmatic engineering sense. Obviously, in reality, the costs should be defined in a more sophisticated way and possibly they should be *different* for *different* levels of *different* actions towards *different* criteria.

Table 1. Available actions and coefficients of influences of the actions on different subcriteria

No.	Action description
act1	Reduce number of workers
act2	Reduce nominal power production
act3	Reduce number of production units
act4	Reduce percentage of contribution to the welfare
act5	Increase number of installed backup components
act6	Increase external emergency measures
act7	Increase duration of backup component
act8	Reduce duration of repair and recovery actions
act9	Enhance training
act10	Enhance safety management
act11	Reduce number of accesses
act12	Enhance entrance control
act13	Enhance surveillance

	Number of workers	Nominal power production	Number of production units	Percentage of contribution to the welfare	Size of served cities	Connection distance	Number of installed backup components	External emergency measures
Actions	Crit1	Crit2	Crit3	Crit4	Crit5	Crit6	Crit7	Crit8
act1	1	0	1	0	0	0	-0.4	-1
act2	0	1	0	1	0	0	0	0
act3	0.7	1	1	0	0	0	0.6	0
act4	0	0	0	1	-1	0	0	0
act5	-0.2	0	0	0	0	0	1	0
act6	-0.1	0	0	0	0	0	0	1
act7	0	0	0	0	0	0	0	0
act8	0	0	0	0	0	0	0	0
act9	0	0	0	0	0	0	0	0
act10	0	0	0	0	0	0	0	0
act11	0	0	0	0	0	0	0	0
act12	0	0	0	0	0	0	0	0
act13	0	0	0	0	0	0	0	0

	duration of backup component	Duration of repair and recovery actions	Training	Safety management	Physical size of the system	Number of accesses	Entrance control	Surveillance
Actions	Crit9	Crit10	Crit11	Crit12	Crit13	Crit14	Crit15	Crit16
act1	0	-0,2	0,5	0,5	0	1	0,4	0,4
act2	0	0	0	0,2	0	0	0	0
act3	0,2	0,3	0,4	0,2	0,7	0	0	0,3
act4	0	0	0	0,1	0	0	0	0
act5	0,5	0	-0,2	0,1	-0,3	0	0	-0,15
act6	0,3	0	-0,1	0,05	0	0	0	-0,05
act7	1	0	0	0,1	0	0	0	0
act8	0	1	-0,2	0,1	0	0	0	0
act9	0	0,5	1	0,2	0	0	0,2	0
act10	0	0	-0,2	1	0	0	0	0
act11	0	0	-0,1	0,1	0	1	0,4	0,1
act12	0	0	-0,1	0,1	0	0	1	0
act13	0	0	0	0,2	0	0	0	1

No.	level1	level2	level3
act1	20%	25%	30%
act2	20%	30%	40%
act3	1	2	3
act4	10%	20%	30%
act5	1	2	3
act6	0,5	1	2
act7	12	24	48
act8	6	12	24
act9	1	3	5
act10	1	3	5
act11	1	2	3
act12	1	2	3
act13	1	2	3

In what follows, two analyses are performed: first, based on the indicators of Section 4, different combinations of actions are ranked according to their ability in reducing the vulnerability of a group of NPPs (Section 5.1); then, the inverse classification problem of Section 3 is tackled using the sensitivity indicators of Section 4 and taking into account the action costs and budget limitations (Section 5.2).

5.1 Ranking different combinations of actions based on $\overline{\Delta M}$

A set *G* of *N* (*N=20*) NPPs ($G = \{NPP_i, i \in \{1,2,\dots,N\}\}$) is available: 10 of them (NPPs from No.6 to No.15 $G^{ref} = \{NPP_i, i \in \{6,7,\dots,15\}\}$) are selected as a reference set to evaluate the sensitivity indicators; the remaining NPPs are regrouped to form a set G^{test}($G^{test} = \{NPP_i, i \in \{1,2,\dots,5\} \cup \{16,17,\dots,N\}\}$) used to test the combinations of actions ranked using G^{ref}. Based on the reference set, we have performed an exhaustive calculation of the value of $\overline{\Delta M}$ for all the possible combinations of actions (in total, 4^13=67108864 combinations). Then, we selected the ones (in total 29940 combinations) that have the (same) highest value of $\overline{\Delta M}$ (i.e., $\overline{\Delta M}$=14): these represent the optimal combinations of actions according to $\overline{\Delta M}$: in what follows, this set is referred to as $Combination_{\overline{\Delta M}}^{highest}$.

All the combinations of actions belonging to the set $Combination_{\overline{\Delta M}}^{highest}$ are applied to each of the *N* (*N=20*) NPPs in *G*: the resulting categories ($C_i^{\lambda\prime}, i \in \{1,2,\dots,N\}$) are reported in Table 2. Note that the actions are ranked according to values of $\overline{\Delta M}$ that are evaluated on a *group* of *reference plants* (G^{ref}): in this view, they provide an indication only on the *expected* performance of the actions on *new* plants and, thus, they may not provide any

indications about the combination of actions that is *optimal* for *one particular* plant. Thus, in order to verify how close these sets of actions are to the combinations that are optimal for a particular NPP, we compare the assignments $C^{\lambda\prime}$ (Table 2) with the best category that each NPP may reach ($C_i^{\lambda}, i \in \{1,2,\dots,N\}$) (in other words, C^{λ} is the category that NPP_i reaches after the application of a combination of actions that is *the optimal one* for *that particular* plant). In order to do so, another exhaustive calculation is done upon the group *G* with the purpose of finding the actions that bring each particular NPP to the best category possible (notice that for some NPPs, reaching category 1 may not be possible). All the possible combinations of actions are tested on *each* NPP in order to find the best assignment C_i^{λ} for each of them. The results are shown in Table 2. The first column of the results shows the original assignments for the NPPs in the studied set *G*. The second column shows the corresponding possibly best assignments C^{λ} and the third column provides the new assignments $C^{\lambda\prime}$ after the application of the combinations of actions included in $Combination_{\overline{\Delta M}}^{highest}$.

Analyzing the best assignments C^{λ} of the NPPs in the reference set ($C_i^{\lambda}, i \in \{6,7,\dots,15\}$), we observe that they coincide perfectly (100%) with the assignments $C^{\lambda\prime}(C_i^{\lambda\prime}, i \in \{6,7,\dots,15\})$ obtained after the application of the actions in $Combination_{\overline{\Delta M}}^{highest}$. If we take the NPPs in the test set as new NPPs and compare the assignments obtained by these two methods with the original assignments $C(C_i, i \in \{1,2,\dots,5\} \cup \{16,17,\dots,N\})$, we find that: (i) all the NPPs are stable or ameliorated after the application of the combinations of actions in $Combination_{\overline{\Delta M}}^{highest}$; (ii) there are 2 out of 10 NPPs that are not ameliorated to the best category C_i^{λ} (i.e., NPPs 16 and 19): they remain in the same category; instead, 8 out of 10 NPPs are ameliorated to their best possible categories: then, the probability that the combinations of actions $Combination_{\overline{\Delta M}}^{highest}$ ameliorate a new NPP to its best possible category C^{λ} is 80%.

Table 2. Comparison of assignments: Best possible Assignment C_i^{λ} and After action

Assignment $C_i^{\lambda'}$ listed with NPPs that are differently assigned highlighted (NPP16, NPP19)

No.	Original Assignment	Best possible Assignment C_i^{λ}	After action Assignment $C_i^{\lambda'}$
NPP1	1	1	1
NPP2	3	3	3
NPP3	2	2	2
NPP4	3	1	1
NPP5	3	2	2
NPP6	2	1	1
NPP7	2	1	1
NPP8	4	2	2
NPP9	4	2	2
NPP10	4	3	3
NPP11	1	1	1
NPP12	2	1	1
NPP13	3	2	2
NPP14	3	1	1
NPP15	4	1	1
NPP16	3	2	3
NPP17	2	2	2
NPP18	3	2	2
NPP19	3	2	3
NPP20	2	1	1

5.2 Constrained inverse problem: identification of the best combination of actions considering constraints

In a more realistic case, the cost of the protective actions should be considered. Although in reality the costs of different actions can be different, and the same action may cost differently when applied to different NPPs, for simplicity, in this paper we define the $Cost$ of a combination of actions (in arbitrary units) as the sum of the levels l_i^j of the actions:

$$Cost = \sum_{i=1}^{N} Cost_i;$$

$$Cost_i = \sum_{j=1}^{F} l_i^j * A_i^j;$$

$$A_i = 1, if\ l_i^j \neq 0;$$

$$A_i = 0, if\ l_i^j = 0;$$

$$i \in \{1,2,\dots,N\}, j \in \{1,2,\dots,F\}. (9)$$

We assume that a budget B_i is allocated for the improvement of the generic power plant NPP_i: the budgets $B_i, i \in \{1,2,\dots,N^{G^{test}}\}$allocated for the NPPs of the test set $G^{test} = \{NPP_i, i \in \{1,2,\dots,5\} \cup \{16,17,\dots,N\}\}$ are shown in Table 3.

Table 3. Budgets available for the NPPs belonging to the test set G^{test}

No.	Budget B_i
NPP1	10
NPP2	25
NPP3	15
NPP4	5
NPP5	16
NPP16	19
NPP17	10
NPP18	23
NPP19	9
NPP20	17

As before, we take the reference set $G^{ref} = \{NPP_i, i \in \{6,7,\dots,15\}\}$ to calculate the value of $\overline{\Delta M}$ for all possible combinations of actions. Then, for each NPP in the test set G^{test}, we identify the combination(s) of actions with the highest value of $\overline{\Delta M}$ and whose costs $Cost_i (i \in \{1,2,\dots,N^{G^{test}}\})$ are lower than or equal to the given budgets B_i:

$$Find\ act_i:$$

$$Max\left(\overline{\Delta M}(NPP_i, act_i)\right);$$

$$act_i \subseteq act;$$

$$s.t.\ Cost_i \leq B_i;$$

$$i \in \{1,2,\dots,N^{G^{test}}\}. \quad (10)$$

The results are shown in Table 4. Among all the possible combinations of actions, the ones that present the highest value of $\overline{\Delta M}$ ($\overline{\Delta M}^{Max} = 14$) have a minimum cost $Cost^{Min} = 19$. So, all the NPPs in the test set G^{test} that have a budget higher than or equal to $Cost^{Min}$ (i.e., NPP2, NPP16 and NPP18) can be ameliorated to their corresponding best possible categories (as presented in Section 4). Five of the remaining NPPs (i.e., NPP1, NPP3, NPP17, NPP19 and NPP20) can still be ameliorated to the same category that would be obtained by the actions in the set $act(with\ Cost \leq 10,15,10,9,17)$, even though they have a budget, which is lower than, $Cost^{Min} = 19$ and a performance lower than $\overline{\Delta M}^{Max} = 14$.

Table 4. Assignments comparison

No.	Original Assignment	Best possible Assignment	Best Assignment	Limited Budget Assignment
NPP1	1	1	1	1
NPP2	3	3	3	3
NPP3	2	2	2	2
NPP4	3	1	1	1 or 2
NPP5	3	2	2	2 or 3
NPP16	3	2	3	3
NPP17	2	2	2	2
NPP18	3	2	2	2
NPP19	3	2	3	3
NPP20	2	1	1	1

The situation is different for NPP4 and NPP5 (Table 5). They are originally assigned to category 3. NPP4 can be ameliorated by any combination of actions belonging to $act(with\ Cost \leq 5)$. Among all the combinations of actions that have the best value of $\overline{\Delta M}$ equal to 7 and cost limited by the given budget, 73.91% can bring NPP4 up to category 2 and 26.09% can bring it to the best category. Instead, NPP5 cannot be ameliorated to the best category by any of the combinations: in particular, 18.52% of the actions leave such NPP in category 3 whereas 81.48% bring it up to category 2.

Table 5. Assignments for NPP4 and NPP5

Assignment	3	2	1
NPP4	0.00%	73.91%	26.09%
NPP5	18.52%	81.48%	0.00%

6 CONCLUSIONS

In this paper, we have developed a pragmatic inverse classification framework for identifying ameliorative action(s) to reduce the vulnerability with respect to intentional hazards of safety-critical systems (in the example of reference, Nuclear Power Plants-NPPs). An MR-sort classification model calibrated on a small-sized set of data representing a priori-known classification examples has been used. Sensitivity indicators have been introduced to evaluate combinations of actions with respect to their ability to reduce the vulnerability of the safety-critical systems considered. A case study referring to NPPs vulnerability to intentional attacks has been worked out. The results show that the actions ranked as best according to the proposed indicators give a satisfactory performance in terms of reduction of vulnerability in

test NPPs, even in presence of budget constraints: for example, in the case without budget constraints eight out of ten NPPs are ameliorated to their best possible categories, whereas two of them remain in the same categories; in the constrained case still six of the ten NPPs are brought to their best possible vulnerability classes.

The proposed methodological framework provides a powerful tool for systematically and pragmatically evaluating the safety and vulnerability as well as other characteristics of critical systems.

For future research, the following issues will be considered. Since one set of weights is usually an insufficient basis for giving priorities, the sensitivity of investment priorities to the weights of criteria can be tackle: for example, in [22][23][24] a "scenario" is introduced that reflects a set of weights for each stakeholder, such as emphasis on particular aspects of safety in the aftermath of a major nuclear incident.

As presented in [25], an influential set of weights can suggest R&D priorities in protection of energy systems.

Moreover, a set of weights can also be brought by other stakeholders, such as owners, operators and users etc: each set of weights presumably leads to variation in the preferred safety investments [26].

In addition, although in this work significant efforts have been made to assign numerical values to the impacts of actions (in order to represent the problem as realistically as possible), in a non-fictitious situation the task is expected to be more complex. Actually, the relations between the actions and the criteria taking into account the dependencies of different attributes and systems are always difficult to identify: in such cases, resorting to the judgment of real experts and possibly to real historical data will be mandatory.

Finally, the inverse classification problem could be tackled within an optimization framework. Proper optimization algorithms could be considered for the optimal selection of protective actions to apply to each considered safety-critical systems (e.g., NPP). The results can, then, be compared with the ones obtained by the sensitivity indicators proposed in the present paper.

REFERENCES

1 Kröger W, Zio E. Vulnerable Systems. UK, London: Springer, 2001.

2 Aven T. Foundations of Risk Analysis. Berlin: Wiley, N.J, 2003.

3 Aven T. Some reflections on uncertainty analysis and management. Reliability Engineering and System Safety, 2010; 95, 195-201.

4 Aven T. Misconceptions of Risk. Chichester: Wiley, 2010.

5 Aven T, Heide B. Reliability and validity of risk analysis. Reliability Engineering and System Safety, 2009; 94, 1862–1868.

6 Wang TR, Mousseau V, Zio E. A Hierarchical Decision Making Framework for Vulnerability Analysis. Proceedings of ESREL2013, Sep 2013, Amsterdam, Netherlands. pp.1-8.

7 Leroy A, Mousseau V, Pirlot M. Learning the parameters of a multiple criteria sorting method, The Second International Conference on Algorithmic Decision Theory,Algorithmic Decision Theory, R.I. Brafman, F. Roberts, and A. Tsoukiàs (Eds.): ADT 2011, LNAI 6992, pp. 219–233, Germany, Berlin: Springer, 2011

8 Charu C, Chen C, Jiawei H. The Inverse Classification Problem, JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(3): 458–468 May 2010

9 Aggarwal CC, Chen C, Jiawei H. On the Inverse Classification Problem and its Applications. Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference, 2006. IEEE.

10 Aiguo L, Xin Z, Jiulong Z. Performance Analysis of Quantitative Attributes Inverse Classification Problem. JOURNAL OF COMPUTERS, Vol.7, No. 5, May 2012.

11. Aven T, Flage R. Use of decision criteria based on expected values to support decision-making in a production assurance and safety setting. Reliability Engineering and System Safety, 2009; 94, 1491-1498.

12 Hofmann M, Kjølle G, Gjerde O. Development of indicators to monitor vulnerabilities in power systems, 2012 International Conference on Probabilistic Safety Assessment and

Management (PSAM 11) & European Safety and RELiability Conference (ESREL 2012); Helsinki, Finland.

13 NWRA, N. W. R. A. Risk assessment methods for water infrastructure systems, Rhode Island Water Resources Center, University of Rhode Island, Kingston, RI. 2012.

14. Roy B. The outranking approach and the foundations of ELECTRE methods. Theory and Decision 31, 1991, 49- 73.

15. Mousseau V., Slowinski R. Inferring an ELECTRE TRI Model from Assignment Examples. Journal of Global Optimization, vol. 12, 1998, 157-174.

16 Wang TR, Mousseau V, Pedroni N, Zio E. Assessing the confidence of a classification-based vulnerability analysis model, Risk Analysis, doi:10.1111/risa. 12305, 2014.

17 Saltelli A, Ratto M, Andres T, Campolongo F, Carboni J, Gatelli D, Saisana M, Tarantola S. Global sensitivity analysis. The primer. Chichester: Wiley, 2008.

18 Iooss B, Lemaître P. A review on global sensitivity analysis methods. Global sensitivity analysis, 2014

19 Saltelli A, Tarantola S, Campolongo F, Ratto M. Sensitivity analysis in practice. Chichester: Wiley, 2004

20 Rocco C, Zio E. Bootstrap-based techiniques for computing confidence intervals in monte carlo system reliability evaluation. Reliability and Maintainability Symposium, 2005. Proceedings. Annual. Page(s): 303 - 307.

21 Levitin, G., Hausken, K., and Ben Haim, H. (2013), "Defending Majority Voting Systems Against a Strategic Attacker," Reliability Engineering & System Safety 111, 1, 37-44.

22 Thekdi, S.A., and J.H. Lambert 2014. Quantification of scenarios and stakeholders influencing priorities for risk mitigation in infrastructure systems. ASCE Journal of Management in Engineering. 30(1):32-40.

23 Karvetski, C.W., and J.H. Lambert 2012. Evaluating deep uncertainties in strategic priority-setting with an application to facility energy investments. Systems Engineering. 15(4): 483-493.

24 Martinez, L.J., J.H. Lambert, and C. Karvetski 2011. Scenario-informed multiple criteria analysis for prioritizing investments in electricity capacity expansion. Reliability Engineering and System Safety. 96: 883-891.

25 Hamilton, M.C., J.H. Lambert, J.W. Keisler, I. Linkov, and F.M. Holcomb. 2013. Research and development priorities for energy islanding of military and industrial installations. ASCE Journal of Infrastructure Systems. 19(3):297-305.

26 Rogerson, E.C. and J.H. Lambert 2012. Prioritizing risk via several expert perspectives with application to airport runway safety. Reliability Engineering and System Safety. 103: 22-34.

27 Joshi, N.N. and J.H. Lambert 2011. Diversification of engineering infrastructure investments for emergent and unknown non-systematic risks. Journal of Risk Research. 14(4): 1466-4461.

28 Lambert, J.H. and M.W. Farrington 2007. Cost-benefit functions for the allocation of security sensors for air contaminants. Reliability Engineering and System Safety. 92(7):930-946.

29 Zio E. An Introduction to the Basics of Reliability and Risk Analysis. World Scientific Publishing Co, 2007.

30 Larsson, J.E. Knowledge-based methods for control systems. PhD dissertation, Lund Institute of Technology, Department of Automatic Control, 1992.

31 Doumpos, M. and C. Zopounidis 2002. Multricriteria Decision Aid Classification Methods, Kluwer Academic Publishers, Netherlands. 2002, ISBN 1- 4020-0805-8.

APPENDIX A CLASSIFICATION MODEL FOR VULNERABILITY ANALYSIS: THE MAJORITY RULE SORTING (MR-SORT) METHOD

The Majority Rule Sorting Model (MR-Sort) method is a simplified version of ELECTRE Tri, an outranking sorting procedure in which the assignment of an alternative to a given category is determined using a complex concordance non-discordance rule (14)(15). We assume that the alternative to be classified (in this paper, a safety- critical system or infrastructure of interests, e.g., a nuclear power plant) can be described by an n-tuple of elements $x = \{x_1, x_2, ..., x_i, ..., x_n\}$, which represent the evaluation of the alternative with respect to a set of n criteria (by way of example, in the present paper the criteria used to evaluate the vulnerability of a safety critical system of interest may include its physical characteristics, social criticality, level of protection and so on: see Section 2). We denote the set of criteria by N = {1, 2, ..., i, ..., n} and assume that the values x_i of criterion i range in the set X_i (20) (for example, in the present paper all the criteria range in [0, 1]). The MR-Sort procedure allows assigning any alternative $x = \{x_1, x_2, \dots, x_i, \dots, x_n \in X = X_1 \times X_2 \times \dots \times X_i \times \dots \times X_n\}$ to a particular pre-defined category (in this paper, a class of vulnerability), in a given ordered set of categories, $\{C^h : h = 1, 2, ..., M\}$; as mentioned in Section 2, $M = 4$ categories are considered in this work: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, A^4 = serious.

To this aim, the model is further specialized in the following way:

- We assume that X_i is a subset of $\mathbb{R}$ for all $i \in \mathbb{N}$ and the sub-intervals $(X_i^1, X_i^2, \dots, X_i^h, \dots, X_i^M)$ of X_i are compatible with the order on the real numbers, i.e., for all $x_i^1 \in X_i^1, x_i^2 \in X_i^2, \dots, x_i^h \in X_i^h, \dots, x_i^M \in X_i^M$, we have $x_i^1 > x_i^2 > \dots > x_i^h > \dots > x_i^M$. We assume furthermore that each interval $X_i^h, h = 1,2, \dots, M-1$ has a smallest element b_i^h, which implies that $x_i^h \geq b_i^h > x_i^{h+1}$. The vector $b^h = \{b_1^h, b_2^h, \dots, b_i^h, \dots, b_n^h\}$ (containing the lower bounds of in the intervals X_i^h of criteria $i = 1,2, \dots, n$ in correspondence of category h) represents the lower limit profile of category C^h.
- There is a weight ω_i associated with each criterion $i = 1,2, \dots, n$, quantifying the

relative importance of criterion *i* in the vulnerability assessment process; notice that the weights are normalized such that $\sum_{i=1}^{n} \omega_i = 1$.

In this framework, a given alternative $x = \{x_1, x_2, \dots, x_i, \dots, x_n\}$ is assigned to category $C^h, h = 2,3, \dots, M-1$, iff

$$\sum_{i \in \mathbb{N}: x_i \geq b_i^h} \omega_i \geq \lambda \ and \ \sum_{i \in \mathbb{N}: x_i \geq b_i^{h-1}} \omega_i < \lambda, \text{(A.1)}$$

where λ is a threshold ($0 \leq \lambda \leq 1$, e.g., in this paper λ=0.9) chosen by the analyst. Rule (A.1) is interpreted as follows. An alternative *x* belongs to category C^h if: 1) its evaluations in correspondence of the n criteria (i.e., the values $\{x_1, x_2, \dots, x_i, \dots, x_n\}$) are at least as good as b_i^h (lower limit of category Ah with respect to criterion *i*), $i = 1,2, \dots, n$, on a subset of criteria that has sufficient importance (in other words, on a subset of criteria that has a weight larger than or equal to the threshold *λ* chosen by the analyst); and at the same time 2) the weight of the subset of criteria on which the evaluations $\{x_1, x_2, \dots, x_i, \dots, x_n\}$ are at least as good as b_i^{h-1} (lower limit of the successive category C^{h-1} with respect to criterion *i*), $i = 1,2, \dots, n$, is not sufficient to justify the assignment of *x* to the successive category C^{h-1}.

Notice that alternative *x* is assigned to the best category C^1 if $\sum_{i \in \mathbb{N}: x_i \geq b_i^1} \omega_i \geq \lambda$ and it is assigned to the worst category C^M if $\sum_{i \in \mathbb{N}: x_i \geq b_i^{M-1}} \omega_i < \lambda$. Finally, it is straightforward to notice that the parameters of such a model are the *(M-1)* · *n* lower limit profiles (*n* limits for each of the *M-1* categories), the n weights of the criteria $\omega_1, \omega_2, \dots, \omega_i, \dots, \omega_n$, and the threshold *λ*, for a total of (*n*·*M+1)* parameters.

Paper (v)

Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent intentional acts: an optimization-based decision-making approach

T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio.

European Journal of Operational Research, 2015, submitted.

Identification of protective actions to reduce the vulnerability of safety-critical systems to malevolent intentional acts: an optimization-based decision-making approach

T. R. WANG

Chair on Systems Science and the Energetic challenge, Foundation EDF,
Ecole Centrale Paris, Grande Voie des Vignes,
F92-295, Chatenay Malabry Cedex

V. MOUSSEAU

Laboratory of Industrial Engineering, Ecole Centrale Paris, Grande Voie des Vignes,
F92-295, Chatenay Malabry Cedex

N. PEDRONI

Chair on Systems Science and the Energetic challenge, Foundation EDF,
Ecole Centrale Paris, Grande Voie des Vignes,
F92-295, Chatenay Malabry Cedex

E. ZIO

Chair on Systems Science and the Energetic challenge, Foundation EDF,
Ecole Centrale Paris, Grande Voie des Vignes,
F92-295, Chatenay Malabry Cedex
Politecnico di Milano, Energy Department, Nuclear Section, c/o Cesnef, via Ponzio 33/A ,
20133, Milan, Italy

ABSTRACT: An empirical classification model based on the Majority Rule Sorting (MR-Sort) method has been previously proposed by the authors to evaluate the vulnerability of safety-critical systems (in particular, nuclear power plants) with respect to malevolent intentional acts. In this paper, the model serves as the basis for an analysis aimed at determining a set of protective actions to be taken (e.g., increasing the number of monitoring devices, reducing the number of accesses to the safety-critical system, etc) in order to effectively reduce the level of vulnerability of the safety-critical systems under consideration.
In particular, the problem is here tackled within an optimization framework: the set of protective actions to implement is chosen as the one minimizing the overall level of vulnerability of a group of safety-critical systems. In this context, three different optimization approaches have been explored: (i) one single classification model is built to evaluate and minimize system vulnerability; (ii) an ensemble of compatible classification models, generated by the bootstrap method, is employed to perform a "robust" optimization, taking as reference the "worst-case" scenario over the group of models; (iii) finally, a distribution of classification models, still obtained by bootstrap, is considered to address vulnerability reduction in a "probabilistic" fashion (i.e., by minimizing the "expected" vulnerability of a fleet of systems). The results are presented and compared with reference to a fictitious example considering nuclear power plants as the safety-critical systems of interest.

KEYWORDS: safety-critical system, malevolent intentional attacks, vulnerability analysis, protective actions, Majority Rule Sorting (MR-Sort), classification model, inverse classification problem, optimization-based approach, bootstrap, robust optimization, probabilistic optimization

1 INTRODUCTION

The vulnerability of safety-critical systems, like nuclear power plants, is of great concern, given the multiple and diverse hazards that they are exposed to (e.g., intentional, random, natural etc.) (Kröger & Zio 2011) and the potential large-scale consequences. This justifies the increased attention for analyses aimed at (i) the systematic identification of the sources of system vulnerability, (ii) the qualitative and quantitative assessment of system vulnerability (Aven 2003)(Aven 2010) and (iii) the definition of effective actions of vulnerability reduction.

The issues at stake involve uncertainty given the long time frame, capital intensive investment and large number of stakeholders with different views and preferences, and call for suitable decision analysis (DA) methods (Leroy, Mousseau, & Pirlot 2001) and particularly multiple criteria decision-making (MCDM) (Doumpos & Zopounidis 2002)(Belton & Stewart 2002). In a previous work (Wang, Mousseau, & Zio 2013), the authors have proposed an empirical classification framework to tackle issues (i) and (ii) above, considering the analysis of the vulnerability of nuclear power plants to malevolent intentional acts. Specifically, we have developed a classification model based on the Majority Rule Sorting (MR-Sort) method (Leroy, Mousseau, & Pirlot 2001) to assign an alternative (i.e., a nuclear power plant) to a given (vulnerability) class (or category). The MR-Sort classification model contains a group of (adjustable) parameters that are calibrated by means of a set of empirical classification examples (also called training set), i.e., a set of alternatives with pre-assigned vulnerability classes (Leroy, Mousseau, & Pirlot 2001)(Wang, Mousseau, & Zio 2013). The performance of the classification-based vulnerability analysis model in terms of accuracy and confidence in the assignments has been thoroughly and systematically assessed in (Wang, Mousseau, Pedroni, & Zio 2014).

In this paper, we are still concerned with intentional hazards (i.e., those related to malevolent acts) and address issue (iii) above, i.e., the definition of the actions to undertake for reducing the level of system vulnerability. In particular, the empirical classification model developed in (Wang, Mousseau, & Zio 2013) is tailored to address the corresponding *inverse (classification)* problem (Aggarwal, Chen, & Han 2010)(Aggarwal, Chen, & Han 2006)(Li, Zhou, & Zhang 2012)(Mousseau & Slowinski 1998), i.e., the problem of determining a set of protective actions which can effectively reduce the vulnerability class of (a group of) safety-critical systems (Aven & Flage 2009), taking into account a specified set of constraints (e.g., budget limits) (Aggarwal, Chen, & Han 2010). Mathematically speaking, the aim is to identify how to modify some features of the input patterns to the classification model (i.e., the attributes of the safety-critical system under analysis) such that the resulting class is changed as desired (i.e., the vulnerability category is reduced to a desired level).

To this aim, an *optimization-based* framework is here undertaken in order to find *one* set of protective actions for *each* of the considered alternatives, such that the *overall* vulnerability level of the *group* of safety-critical systems under consideration is *minimized* under given constraints. In this context, three different optimization approaches have been sought: (i) *one single* classification model is built to evaluate and minimize system vulnerability, (ii) an *ensemble* of

compatible classification models, generated by the bootstrap method, is employed to perform a "robust" optimization, taking as reference the "worst-case" scenario over the group of models; (iii) finally, a distribution of classification models, still obtained by bootstrap, is considered for the vulnerability reduction task, by minimizing the "expected" vulnerability of the fleet of plants. All three optimization problems are numerically solved by CPLEX. The remainder of the paper is structured as follows. Section 2 recalls the classification model for the assessment of vulnerability to intentional hazards. With reference to that, Section 3 introduces the problem of inverse classification for choosing protective actions and the optimization decision-making approach. In Section 4, case studies are proposed to show the applications of the method. Finally, Section 5 gives the discussion and analysis of the results. The conclusions of this research is drawn in Section 6.

2 CLASSIFICATION MODEL FOR THE ASSESSMENT OF VULNERABILITY TO INTENTIONAL HAZARDS

We limit the vulnerability analysis of a system to the evaluation of the susceptibility to intentional hazards and adopt the three-layers hierarchical model developed in (Wang, Mousseau, & Zio 2013) (Figure 1). The susceptibility to intentional hazards (layer 1 in Figure 1) is characterized in terms of attractiveness and accessibility (layer 2 in Figure 1). These attributes are hierarchically broken down into factors which influence them, including resilience interpreted as pre-attack protection (which influences on accessibility) and post-attack recovery (which influences on attractiveness). The disaggregation is made in n criteria (layer 3 in Figure 1) described by the n-tuple $MCrit = \{MCrit_1, MCrit_2, ..., MCrit_i, ..., MCrit_n\}$ with $n = 6$ in this case: physical characteristics ($MCrit_1$), social criticality ($MCrit_2$), possibility of cascading failures ($MCrit_3$), recovery means ($MCrit_4$), human preparedness ($MCrit_5$) and level of protection ($MCrit_6$). These six criteria are further decomposed into a layer of $m = 16$ basic subcriteria $\{crit^j, j = 1, 2, ..., m = 16\}$(layer 4 in Figure 1), for which data and information are collected in terms of quantitative values or linguistic terms depending on the nature of the subcriterion. The descriptive terms and/ or values of the fourth layer subcriteria are, then, scaled to numerical categories. Finally, to get the value of the six third-layer criteria $MCrit = \{MCrit_1, MCrit_2, ..., MCrit_i, ..., MCrit_n\}, n = 6$, (i) we assign weights to each subcriterion to indicate their importance and (ii) we apply a simple weighted sum to the categorical values of the constituent subcriteria $\{crit^j = j = 1, 2, ..., m = 16\}$. These $m = 16$ criteria

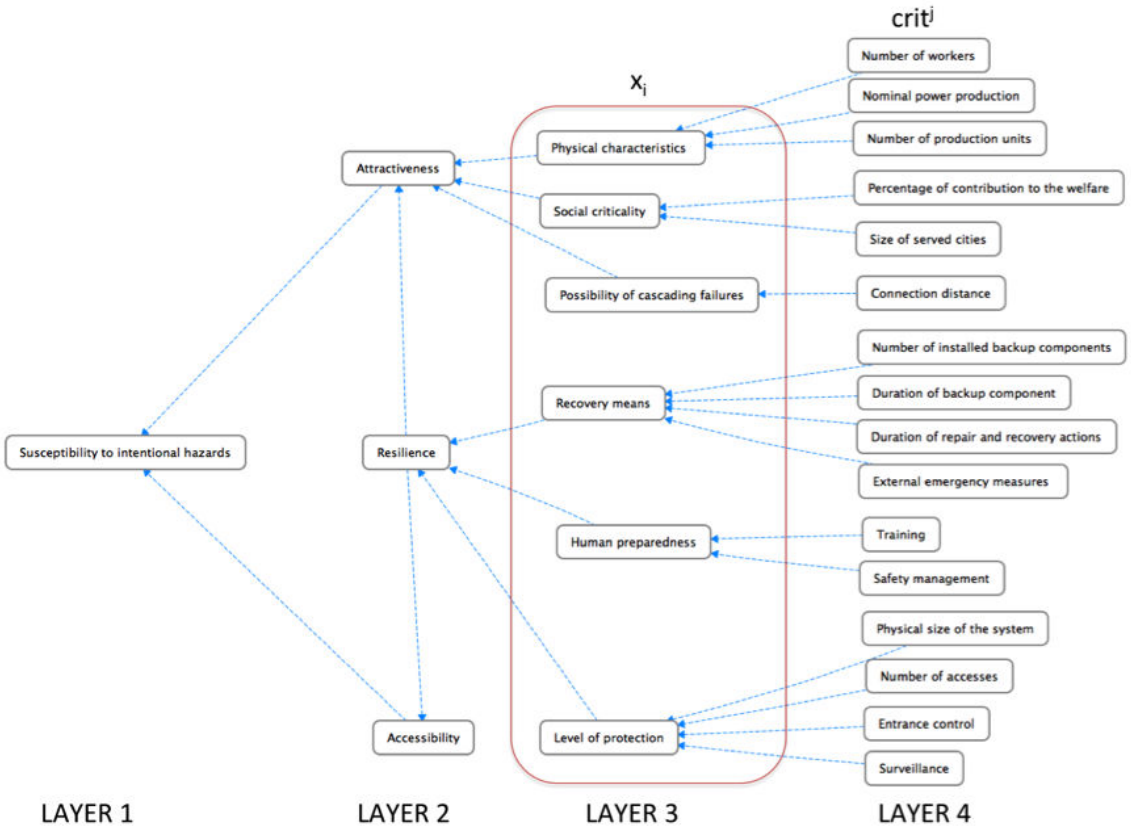

Figure 1: Hierarchical model for susceptibility to intentional hazards

$\{crit^j = j = 1, 2, ..., m = 16\}$ are evaluated to assess the vulnerability of a given safety-critical system of interest (e.g., a nuclear power plant NPP). For the purpose of the present analysis, $M = 4$ levels (or categories) of system vulnerability $\{A^h : h = 1, 2, 3, 4\}$ are considered: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic, A^4 = serious. Then, the assessment of vulnerability corresponds to a classification problem: given the definition of the characteristics of a critical system in terms of the sixteen criteria above, assign the vulnerability category (or class) to which the system belongs.

The classification model, indicated as $M(\cdot|(\omega, b))$, is based on the Majority Rule Sorting (MR-Sort) method (Leroy, Mousseau, & Pirlot 2001)(Roy 1991)(Mousseau & Slowinski 1998). The model contains a group of (adjustable) parameters, i.e., the weights $\omega = \{\omega_1, \omega_2, ..., \omega_n\}$ of the n criteria and the lower "boundaries" defining each category $h = 1, 2, , k$ with respect to each criterion $i = 1, 2, ...n$ (notice that $b^h = \{b_1^h, b_2^h, ..., b_i^h, ..., b_n^h\}$): further details can be found in (Wang, Mousseau, Pedroni, & Zio 2014). These parameters are calibrated through a disaggregation process by means of a set of empirical classification examples (the training set $D_{TR} = \{(x_p, \Gamma_p^t), p = 1, 2, ..., N\}$, i.e., a set of N alternatives $x_p = \{x_1^p, x_2^p, ..., x_i^p, ..., x_n^p\}, p = 1, 2, ..., N$ together with the corresponding real pre-assigned categories (i.e., vulnerability classes) Γ_p^t (the

superscript t indicates that Γ_p^t represents the true, a priori-known vulnerability class of alternative x_p). Further details about the generation of classification models are not reported here for brevity: the interested reader is referred to (Wang, Mousseau, Pedroni, & Zio 2014).

3 INVERSE CLASSIFICATION PROBLEM FOR PROTECTIVE ACTIONS IDENTIFICATION: AN OPTIMIZATION-BASED DECISION MAKING APPROACH

We define an inverse classification problem aimed at finding a combination of actions reducing the vulnerability of a (group of) safety-critical system(s) eventually under budget limitations.
To illustrate the methodology, we consider a set of N alternatives ($x_p, p \in \{1, 2, ..., N\}$) characterized by $m = 16$ basic features ($crit^j, j \in \{1, 2, ..., m\}$). Each vector x_p represents one safety-critical system (in our case, a NPP). On the basis of these $m = 16$ features, the NPPs are assigned to $M = 4$ pre-defined categories ($\{A^h : h = 1, 2, 3, 4\}$), where A^1 represents the best situation, i.e., lowest vulnerability, as presented in the previous section. Let $act = \{act^1, act^2, ..., act^F\}$ denote the available set of actions, each of which can influence on one or more basic criteria $crit^j, j \in \{1, 2, ..., m\}$ (Figure 2) with different intensity, as measured by a set of coefficients $coeff^{kj}, k \in \{1, 2, ..., F\}, j \in \{1, 2, ..., m\}$. In other words, $coeff^{kj}$ is the "weight" of the influence of action k on attribute j (the higher the absolute value of $coeff^{kj}$, the stronger the effect of action k on attribute j). Notice that a positive (resp. negative) coefficient $coeff^{kj}$ means that action k has an ameliorative (resp. deteriorative) effect on attribute j, whereas if $coeff^{kj}$ is equal to zero, then criterion j is not influenced by action k. The implementation of one or more actions modifies the attribute values $crit^j, j \in \{1, 2, ..., m\}$ and as a result, the vulnerability of the system (i.e., the assignment by the classification model) may change. In this paper, we assume that the total effect of the available set of actions $act = \{act^1, act^2, ..., act^F\}$ on criterion j is obtained by a linear superposition of the effects of each action act^k:

$$crit'^j = crit^j + \sum_{k=1}^{F} coeff^{kj} * act^k, k \in \{1, 2, ..., F\}, j \in \{1, 2, ..., m\}. \quad (1)$$

where $crit'^j$ is the value of attribute j after the identified set of available actions has been implemented.

Also, let $Cost(x_p, act'), act' \subseteq act$ denote the cost of the combination of actions act' applied to x_p. If $c_k^p (p \in \{1, 2, ..., N\}, k \in \{1, 2, ..., F\})$ is the cost of action k on x_p, then:

$$Cost(x_p, act') = \sum_k c_k^p, k \in \{1, 2, ..., F\}. \quad (2)$$

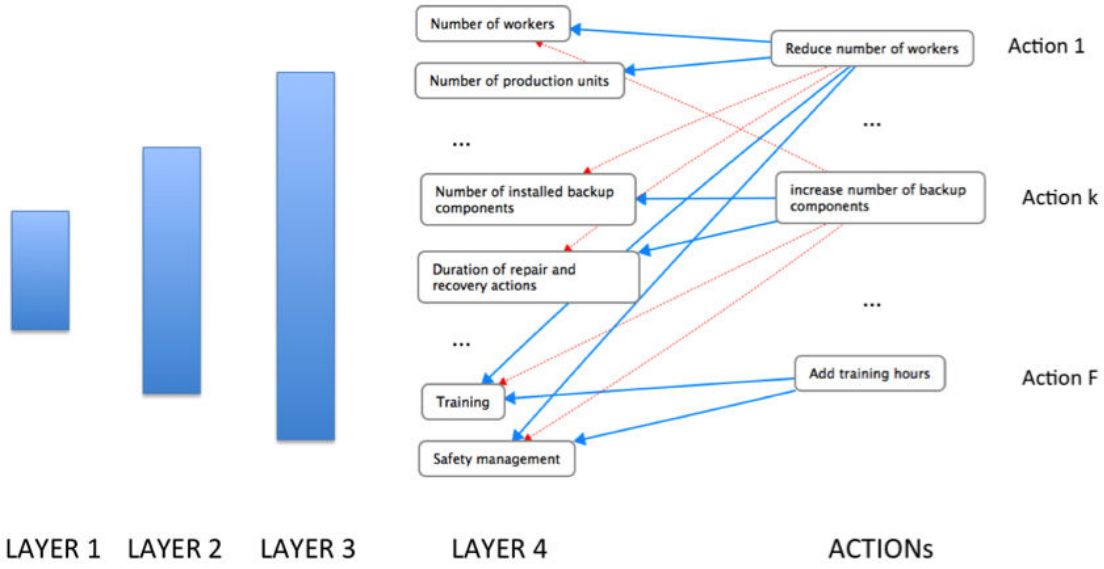

Figure 2: Schema of direct actions for basic criteria

The inverse classification problem can, then, be formulated as follows: given a limited budget B_g for the entire group of NPPs considered, find out for each of the NPPs the best combination of actions that provide the maximal possible reduction in their vulnerability level $A_p^{\lambda'}, A_p^{\lambda'} \in \{1,2,3,4\}, p \in \{1,2,...,N\}$(as presented in the previous Section, the smaller the category value, the less vulnerable the NPP). In particular, we have chosen the strategy to reduce, under budget constraint, the global vulnerability of a group of alternatives in giving priority to the NPPs that are originally assigned to the worst category; in other words, we try to maximize a properly weighted sum of the ameliorations in the vulnerability categories undergone by all the NPPs. This is mathematically represented by the objective function:

$$I^x = \rho_3 * Q_{43} + \rho_2 * Q_{32} + \rho_1 * Q_{21} \tag{3}$$

where $Q_{n(n-1)}(n \in \mathbb{Z})$ represents the number of NPPs among the N available ones $\{x|x_p, p \in \{1,2,...,N\}\}$ that are ameliorated from category A^n to category A^{n-1} by a given combination of actions. The constants $\{\rho_i|i \in \{1,2,3\}\}$ represent weights that we assign to the number of ameliorated NPPs $Q_{n(n-1)}(n \in \mathbb{Z})$, in particular:

$$\rho_3 = 100, \rho_2 = 50, \rho_1 = 25. \tag{4}$$

In this case, by maximizing the objective function I^x high importance is given to the amelioration of the worst (i.e., most vulnerable) NPPs.

In this context, three different optimization approaches have been undertaken: (i) one single classification model is built to evaluate and minimize system vulnerability, (ii) an ensemble of

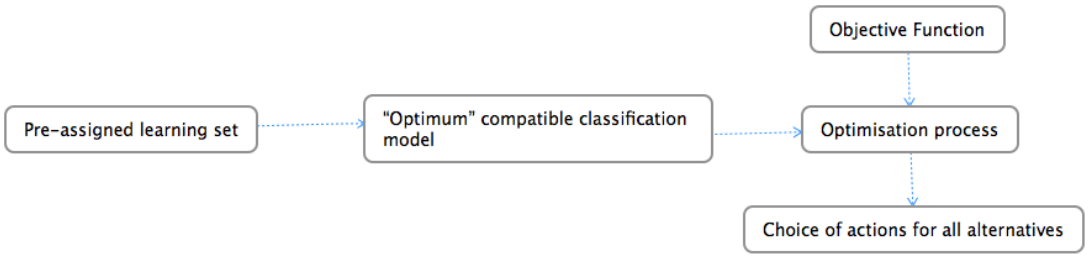

Figure 3: Representation of Simple Optimization

compatible classification models, generated by the bootstrap method, is employed to perform a "robust" optimization, by considering the "worst-case" scenario; (iii) finally, a distribution of classification models, still obtained by bootstrap, is considered to address vulnerability reduction in a "probabilistic" fashion.

3.1 *Simple Optimization*

As presented in Section 2, and in more details in (Wang, Mousseau, Pedroni, & Zio 2014), we can construct a classification model as $M^*(\cdot|\omega^*, b^*)$(with ω^* the weights and b^* the lower profiles) compatible with all the pre-assigned alternatives in the training set D_{TR} through a disaggregation process. We name this model the "optimum" classification model. The optimization-based inverse classification process aims at finding an optimal set of actions for each of the NPPs for which the objective function I^x is maximized: this will improve the performance of the group of NPPs, while giving priority to the worst ones. In more detail, the problem can be formulated as follows:

$$Find\ act'_p = arg\ Max_{\{act'_p, p=1,2,...,N\}}(I^x(act'_p, M^*)), \tag{5}$$

$$s.t.\ \sum_p Cost(x_p, act'_p) \leq B_g, \tag{6}$$

$$\{x|x_p, p \in \{1, 2, ..., N\}\} \tag{7}$$

Under the constraint of budget limitation, we find the combination of protective actions that maximize the value of the objective function I^x, presented above.

3.2 *Robust Optimization*

The optimization approach introduced above provides a choice of protective actions for the NPPs using (only) the "*optimum*" classification model $M^*(\cdot|\omega^*, b^*)$. However, for the training

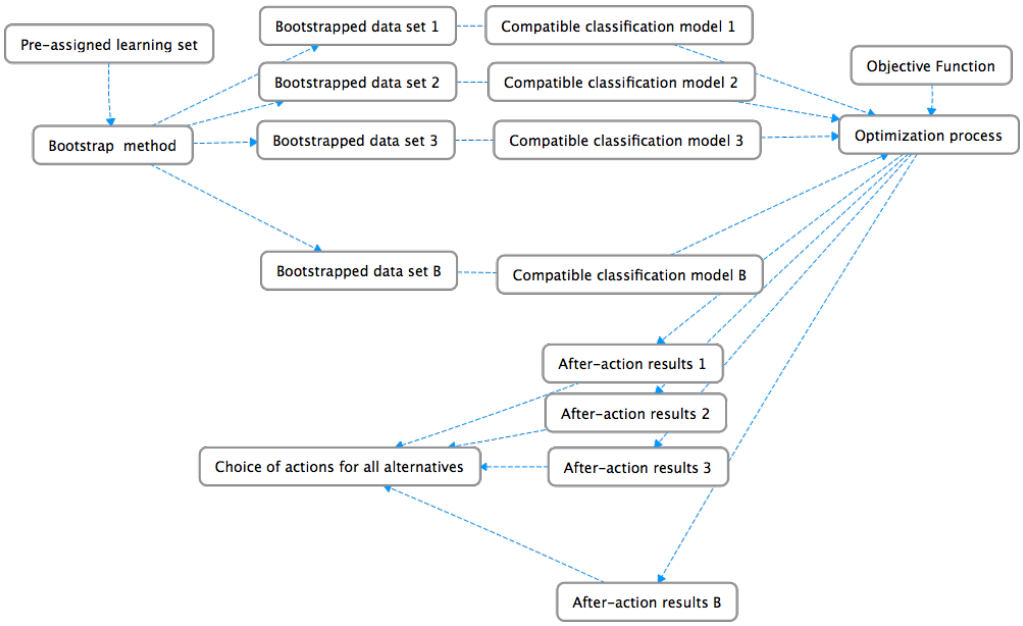

Figure 4: Representation of Robust Optimization

set of pre-assigned alternatives there are a number of compatible classification models. To account for this model uncertainty, we aim at finding the combination of protective actions (for each of the NPPs) that can ameliorate the NPPs to a satisfactorily low level of vulnerability, considering all compatible classification models. In other words, the combination of actions that we obtain should be "*robust*" to the (model) uncertainty arising from the fact that the empirical classification model is trained with a *finite* set of data and, thus, multiple models are compatible.

To this aim, the bootstrap method (Efron & Thibshirani 1993) is applied to create an ensemble of classification models constructed on different data sets bootstrapped from the original one (Zio 2006). The basic idea is to generate different training datasets by random sampling with replacement from the original one (Efron & Thibshirani 1993): such different training sets are used to build different individual classification models of the ensemble. In this way, the individual classifiers of the ensemble possibly perform well in different regions of the training space. In more detail, the main steps of the bootstrap algorithm are as follows (Figure 4):

a. Generate a bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$, by performing random sampling with replacement from the original data set $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$ of N input/output patterns. The data set $D_{TR,q}$ is thus constituted by the same number N of input/output patterns drawn among those in D_{TR}, although due to the sampling with

replacement some of the patterns in D_{TR} will appear more than once in $D_{TR,q}$, whereas some others will not appear at all.

b. Build a classification model $\{M_q(\cdot|\omega_q, b_q) : q = 1, 2, ..., B\}$, on the basis of the bootstrap data set $D_{TR,q} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N\}$.

Given the bootstrapped ensemble, the mathematical formulation of the robust optimization is as follows:

$$Find\ act'_p = arg\ Max_{\{act'_p, p=1,2,...,N\}}\ Min_q(I^x(act'_p, M_q)), \tag{8}$$

$$s.t.\ \sum_p Cost(x_p, act'_p) \leq B_g, \tag{9}$$

$$\{x|x_p, p \in \{1, 2, ..., N\}\}, \tag{10}$$

$$\{M|M_q \in M, q \in \{1, 2, ..., B\}\}. \tag{11}$$

A large number $B(= 100)$ of compatible classification models $\{M|M_q \in M, q \in \{1, 2, ..., B\}\}$ are typically generated by bootstrap. Correspondingly, the minimum value $Min_M(I^x(act'_p, M_q))$ of objective function $I^x(act'_p, M_q)$ over the B compatible models in correspondence of each set of actions can be gathered. In particular, a distribution of vulnerability classes can be obtained for each NPP. Then, based on the distribution and applying the majority-voting rule, we assign each NPP to its most likely after-action category. Then, the optimization solver aims at finding the optimal combination of actions that robustly and conservatively maximize the worst value of the objective function $I^x(act'_p, M_q)$.

In more detail, the robust optimization algorithm proceeds as follows:

1. The solver proposes a set of actions for each x_p; each bootstrapped classification model $M_q(\cdot|\omega_q, b_q)$ is used to provide an after-action vulnerability class $\Gamma_p^q, q = 1, 2, ..., B$ to each alternative of interest, i.e., $\Gamma_p^q = M_q(x_p|\omega_q, b_q)$;

2. On the basis of the results obtained at step 1 above, a value for function $I^x(act'_p, M_q)$ is computed for *each* compatible model $M_q(\cdot|\omega_q, b_q), q = 1, 2, ..., B$, to obtain an ensemble of values $I^x(act'_p, M_q)$;

3. The minimum (i.e., worst) value among $I^x(act'_p, M_q), q = 1, 2, ..., B$, is taken as the objective function to maximize; in other words, we aim at identifying the set of actions able to improve the "worst-case scenario" over the possible compatible models;

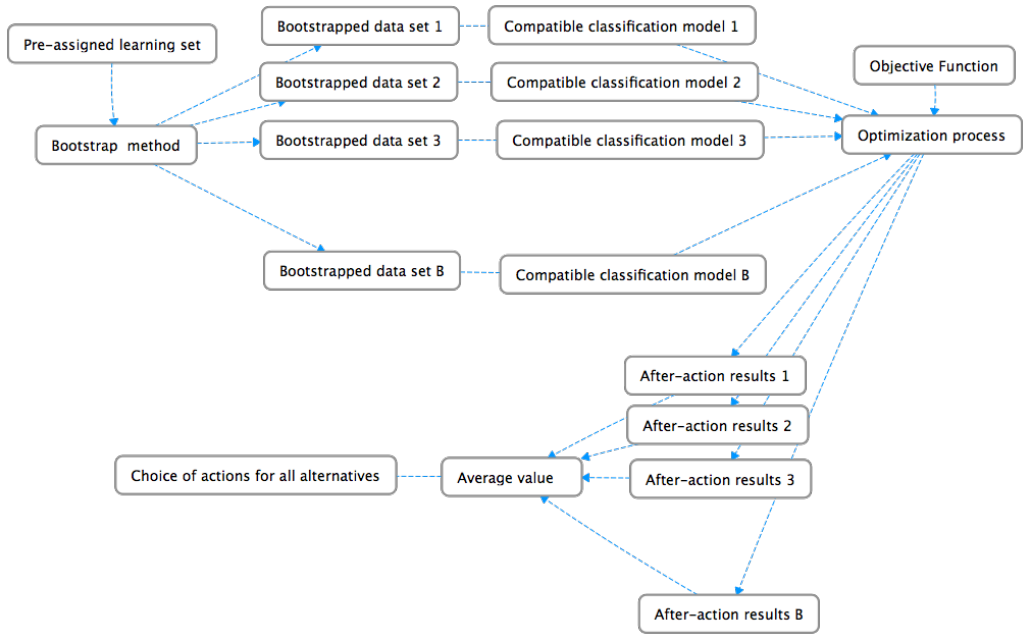

Figure 5: Representation of Probabilistic Optimization

4. We repeat the steps above for different combinations of actions $act'_p, p = 1, 2, ..., N$ in order to find out the combination of actions for each of the considered NPPs that can ameliorate the worst case situation as much as possible.

3.3 *Probabilistic Optimization*

The main steps (Figure. 5) are the same as those of the Robust Optimization presented in Figure 4, but the objective function is changed. Instead of improving the worst case over all the models, we choose to improve the expected value of the probablity distribution of the function I^x. Thus, in this case, we "ignore" some of the "extreme" classification models generated by bootstrap.

The mathematical formulation of the problem is as follows:

$$Find\ act'_p = arg\ Max_{\{act'_p, p=1,2,...,N\}} \frac{1}{B} \sum_{q=1}^{B} (I^x(act'_p, M_q)), \tag{12}$$

$$s.t. \sum_{p} Cost(x_p, act'_p) \leq B_g, \tag{13}$$

$$\{x | x_p, p \in \{1, 2, ..., N\}\}, \tag{14}$$

$$\{M | M_q \in M, q \in \{1, 2, ..., B\}\}. \tag{15}$$

Table 1: Basic Criteria

No.	Basic Criteria
$crit^1$	Number of workers
$crit^2$	Nominal power production
$crit^3$	Number of production units
$crit^4$	Percentage of contribution to the welfare
$crit^5$	Size of served cities
$crit^6$	Connection distance
$crit^7$	Number of installed backup components
$crit^8$	Duration of backup component
$crit^9$	Duration of repair and recovery actions
$crit^{10}$	External emergency measures
$crit^{11}$	Training
$crit^{12}$	Safety management
$crit^{13}$	Physical size of the system
$crit^{14}$	Number of accesses
$crit^{15}$	Entrance control
$crit^{16}$	Surveillance

4 APPLICATION

The methods presented in Section 3 are applied on a case study concerning the vulnerability analysis of NPPs (Mousseau & Slowinski 1998). We identify $n = 6$ main criteria $i = 1, 2, ..., n = 6$ by means of the hierarchical approach presented in (Leroy, Mousseau, & Pirlot 2001) (see Section 2): $MCrit_1$ = physical characteristics, $MCrit_2$ = social criticality, $MCrit_3$ = possibility of cascading failures, $MCrit_4$ = recovery means, $MCrit_5$ = human preparedness and $MCrit_6$ = level of protection. Then, these 6 criteria are decomposed into $m = 16$ basic criteria $\{crit^j, j = 1, 2, ..., m = 16\}$ (see Table 1). Finally, $k = 4$ vulnerability categories $A^h, h = 1, 2, 3, 4$ are defined as: A^1 = satisfactory, A^2 = acceptable, A^3 = problematic and A^4 = serious (Section 2).

The training set D_{TR} is constituted by a group of $N = 18$ NPPs with corresponding a priori-known categories Γ_p^t (all the considered NPPs are pre-assigned to A^2, A^3 *or* A^4, since the alternatives originally assigned to best category A^1 are not taken into account), i.e., $D_{TR} = \{(x_p, \Gamma_p^t) : p = 1, 2, ..., N = 18\}$. The training set of plants with the corresponding values of the basic criteria is summarized in Table 2. Taking as reference the 16 basic criteria, 13 actions "directly" impacting on them are defined (Table 3): for example, for the criterion *Number of workers*, the direct action is *Reduce number of workers*. On the other hand, there are certain

Table 2: Training set with $N = 18$ assigned alternatives

Alternatives, x_p	Vulnerability Class, Γ_p^t
x_1={1600, 3600, 4, 90, 350000, 2, 4, 1, 110, 70, 3, 0, 500, 2, 4, 5}	A^3
x_2={800, 2600, 3, 60, 300000, 10, 2, 3, 90, 25, 7, 3, 300, 2, 4, 4}	A^2
x_3={900, 2600, 3, 70, 250000, 30, 3, 2, 130, 30, 7, 3, 400, 1, 2, 3}	A^3
x_4={1000, 3000, 4, 70, 500000, 12, 2, 1, 60, 20, 3, 3, 400, 1, 3, 4}	A^3
x_5={1400, 3600, 4, 80, 400000, 50, 1, 2, 140, 30, 7, 3, 500, 1, 4, 4}	A^2
x_6={800, 1800, 2, 50, 18000, 10, 1, 2, 85, 10, 7, 7, 250, 2, 3, 3}	A^2
x_7={1600, 5200, 6, 100, 800000, 10, 3, 0, 70, 35, 0, 0, 500, 2, 3, 4}	A^4
x_8={2000, 5400, 6, 100, 1200000, 9, 2, 1, 95, 60, 3, 3, 600, 3, 4, 4}	A^4
x_9={2000, 5400, 6, 100, 1200000, 2, 1, 0, 60, 70, 0, 0, 600, 3, 2, 3}	A^4
x_{10}={830, 2600, 2, 30, 1810994, 70, 2, 3, 135, 15, 7, 3, 280, 1, 5, 5}	A^2
x_{11}={1400, 5200, 4, 100, 100000, 10, 1, 2, 65, 20, 3, 3, 415, 1, 5, 5}	A^3
x_{12}={1000, 1800, 2, 64.2, 556000, 50, 2, 2, 100, 25, 3, 3, 106, 1, 3, 4}	A^3
x_{13}={2200, 5400, 6, 100, 4033197, 35, 2, 3, 130, 55, 15, 7, 152, 1, 5, 4}	A^4
x_{14}={684, 1800, 2, 72, 2538590, 30, 2, 3, 125, 30, 3, 3, 60, 1, 3, 4}	A^3
x_{15}={688, 2600, 2, 100, 2538590, 10, 2, 3, 120, 35, 7, 3, 170, 1, 4, 4}	A^2
x_{16}={2200, 3600, 4, 100, 3258000, 40, 3, 1, 140, 32, 7, 3, 255, 1, 4, 5}	A^3
x_{17}={906, 3000, 2, 100, 1760575, 30, 2, 2, 70, 28, 3, 3, 260, 1, 4, 5}	A^3
x_{18}={1300, 2600, 2, 70, 6272467, 30, 3, 1, 135, 30, 7, 7, 180, 1, 3, 3}	A^2

basic criteria that cannot have a corresponding action: for example, for criterion *connection distance*; it is not possible to physically reduce the distance between sites. Additionally, in our case study, the actions have 3 different influence/impact levels; for example, with reference to action *Reduce number of workers*, the 3 levels imply a reduction of the number of workers of the chosen site by 1) 20%, 2) 25% or 3) 30%. This adds degrees of freedom to the choice of actions. Considering the costs associated to the actions, for the sake of simplicity, we define the cost to be "1" for level 1, "2" for level 2 and "3" for level 3, in relative units, for all actions and NPPs. In what follows, the three optimization-based approaches of Section 3 are applied to obtain the "best" combination of protective actions for each of the NPPs.

4.1 *Simple Optimization*

Two tests are first carried out considering an unlimited or limited budget. The example with unlimited budget aims at showing an ideal case of the inverse classification problem that would lead, in principle, to the best after-action condition. It can be seen that, based on the original dataset of pre-assigned alternatives, there are certain NPPs (i.e., x_1, x_2 *and* x_{15}) that can never be ameliorated to the best category A^1 (Table 4). The identification of the best (i.e., lowest) vulnerability category that one NPP can be assigned to without budget restrictions represents an important base information that provides the decision makers with a global view of the problem

Table 3: Available protective actions

No.	Direct action effect
act^1	Reduce number of workers
act^2	Decrease nominal power production
act^3	Cut down number of production units
act^4	Decrease percentage of contribution to the welfare
act^5	Add installed backup components
act^6	Add external emergency measures
act^7	Prolong the duration of backup components
act^8	Reduce the duration of repair and recovery actions
act^9	Enhance training
act^{10}	Strengthen safety management
act^{11}	Decrease number of accesses
act^{12}	Enhance entrance control
act^{13}	Strengthen surveillance

goals.

The optimization performed with budget constraints aims at solving the realistic problem of finding out the combination of protective actions for each NPP, that ameliorate the group of NPPs with priority to the most vulnerable ones, managing the "residual" resource to improve the others. With an unlimited budget, most of the NPPs are ameliorated to a lower level of vulnerability. Actually, x_1, x_2 *and* x_{15} do not change class because of their particular characteristics (e.g., the physical distance between the site and the nearby cities is closer with respect to that of the other plants, and such characteristics cannot be modified by any action). The minimum

Table 4: After-action assignments of the considered NPPs without budget constraint. White cases in the third column indicate unchanged assignment.

Unlimited budget	Original assignment	Best after-action assignment
x1	3	3
x2	2	2
x3	3	1
x4	3	2
x5	2	1
x6	2	1
x7	4	2
x8	4	2
x9	4	3
x10	2	1
x11	3	2
x12	3	1
x13	4	1
x14	3	2
x15	2	2
x16	3	2
x17	3	2
x18	2	1

Table 5: After-action assignments of the considered NPPs with budget constraint $B_{g_{min}} = 40$ (simple optimization). White cases in the third column indicate unchanged assignment.

Budget constraint: Bg=40	Original assignment	Best after-action assignment
x1	3	3
x2	2	2
x3	3	1
x4	3	2
x5	2	2
x6	2	1
x7	4	2
x8	4	3
x9	4	3
x10	2	2
x11	3	2
x12	3	2
x13	4	2
x14	3	2
x15	2	2
x16	3	2
x17	3	3
x18	2	1

cost necessary to improve each NPP to the best possible category is $B_{g_{min}} = 78$. Fixing a limited budget to $B_{g_{min}} = 40$, the optimization of the actions leads to the ameliorations reported in Table 5. Obviously, x_1, x_2 *and* x_{15} still do not change class as in the case with unlimited budget. Moreover, since the budget is lower than that necessary to ameliorate all NPPs to their best category ($B_{g_{min}} = 78$), there are other NPPs (x_5, x_{10} *and* x_{17}) whose vulnerability category is not changed. On the contrary, all NPPs originally assigned to the "worst" category A^4 improve after action(s); then, the rest of the budget is distributed to ameliorate the other NPPs as much as possible. For example, x_8 *and* x_{12} are improved by one category, whereas they can be improved by two categories in the case of unlimited budget (Table 4).

In the next two subsections, we present the results of the other two optimizations approaches considering only the realistic case of limited budget.

4.2 *Robust Optimization*

The results in the case of limited budget, $B_g = 40$, are shown in Table 6 and compared to the original categories (obtained by majority-voting over the B compatible bootstrapped classification models). There are only 4 NPPs that are ameliorated: x_{13} is ameliorated from A^4 to A^3; x_2, x_3 *and* x_{18} are ameliorated from A^3 to A^2. There are changes in the bootstrapped distributions of the categories of the other NPPs, but not consistent enough to change their final assignments by majority-voting. In comparison with the results obtained in the previous subsection, there are less NPPs that are ameliorated. This is reasonable for a "robust" solution, since

Table 6: After-action assignments of the considered NPPs with budget constraint (robust optimization). White cases in the third column indicate unchanged assignment. MV = majority-voting

Budget constraint: Bg=40	Original assignment by MV	Robust optimization after-action by MV
x1	3	3
x2	3	2
x3	3	2
x4	3	3
x5	3	3
x6	2	2
x7	4	4
x8	4	4
x9	4	4
x10	2	2
x11	3	3
x12	3	3
x13	4	3
x14	3	3
x15	4	4
x16	3	3
x17	4	4
x18	3	2

"extreme" (worst-case) compatible classification models affect the optimization.

4.3 *Probabilistic Optimization*

The probabilistic case is a variation of the Robust Case of Section 4.2. Instead of maximizing $Min_M(I^x(act'_p, M_q))$(i.e., the worst after-action objective function value), we choose to maximize the expected value of the bootstrapped probability distribution of the weighted objective function $I^x(act'_p, M_q)$.

The results are shown in Table 7, in comparison with the original majority-voting category of each NPP. There are 8 NPPs that are ameliorated: $x_8, x_{13}, x_{15}\ and\ x_{17}$ are changed from category A^4 to A^3; $x_2, x_3\ and\ x_{12}$ are changed from A^3 to A^2. In comparison with the results of Section 4.1, there are less NPPs that are ameliorated; in addition, not all the NPPs that were originally assigned to the worst category (A^4) are improved. On the other hand, with respect to the results of the robust optimization (which also considers an ensemble of different compatible models), the group of NPPs is globally improved. The results of the probabilistic case are more satisfactory since most of the NPPs that were assigned to the worst category (A^4) are improved; then, the rest of the resources is used to ameliorate those plants that were assigned to the second worst category (A^3).

Table 7: After-action assignments of the considered NPPs with budget constraint (probabilistic optimization). White cases in the third column indicate unchanged assignment. MV = majority-voting

Budget constraint: Bg=40	Original assignment by MV	Probabilistic optimization after-action by MV
x1	3	3
x2	3	2
x3	3	2
x4	3	3
x5	3	2
x6	2	2
x7	4	4
x8	4	3
x9	4	4
x10	2	2
x11	3	3
x12	3	2
x13	4	3
x14	3	3
x15	4	3
x16	3	3
x17	4	3
x18	3	3

5 DISCUSSION AND ANALYSIS OF THE RESULTS

The three optimizations considered provide conceptually and practically different solutions to the choice of protective actions for the NPPs.

The simple optimization provides a quite specific and limited indication of the amelioration capability of a set of actions with reference to a single classification model with given characteristics. In this case, the classification model is fixed (generated through a disaggregation process based on the number of real-world classification examples available): the number n of main criteria, the number m of basic criteria and the number M of categories (given by the analysts according to the characteristics of the systems at hand). On this basis, the space of all possible combinations of actions for each of the NPPs of the group (and consequently the space of all possible objective functions with the structure mentioned above, i.e., n criteria and M categories) are exhaustively explored by the optimization solver. The weighted objective function defined fulfills the original purpose of ameliorating the NPPs group overall performance, giving preference to those NPPs that were originally assigned to the worst categories.

The robust optimization is inevitably more conservative, given the uncertainty in the compatible models. By the bootstrap method applied on the training set available, an ensemble of B compatible models is built. By so doing, we explore the space of all the classification models compatible with that particular training set. In this view, the bootstrap serves the purpose of accounting for the uncertainty related to using a specific and finite (training) data set for build-

ing a classification model of given structure (i.e., with given numbers n and M of criteria and categories, respectively). In addition, the objective function I^x for the optimization represents the "worst-case scenario" over all models and this injects additional conservatism in the choice of the protective actions for the NPPs.

The probabilistic optimization approach applied to the same set of B compatible models aims at the objective of maximizing the expected value of the weighted function I^x. The overall improvement of the NPPs turns out to be satisfying. Comparisons and thorough discussions are presented in the following subsections.

5.1 *Comparison of the assignments of the NPPs after protective actions*

Three different perspectives of optimization have been carried out under a limited budget ($B_g = 40$), where the simple case considers one "optimum" classification model, whereas the robust and probabilistic cases consider $B(= 100)$ compatible models obtained by the bootstrap method. For fair comparison of the after-action assignments, an adaptation of the results of the simple case is needed.

The set of protective actions generated in the simple case is now applied to the group of alternatives for all B compatible models of the robust and probabilistic cases. Then, the assignments are obtained by the majority-voting rule. This shows the effect that a set of action "optimistically" obtained by resorting to one single "optimum" model has on the NPPs, when applied to an ensemble of compatible models in light of the uncertainties. The results are listed in Table 8. First, we compare the data of the first and second columns. In the first column, there are the original assignments for all the NPPs, evaluated by the single "optimum" classification model; in the second column, there are the assignments for the same group of NPPs obtained by majority voting based on the $B(= 100)$ models. It can be seen that there are some differences of assignments for some NPPs ($x_2, x_5, x_{15}, x_{17}\ and\ x_{18}$): with the single "optimum" model, the vulnerability of these NPPs is "underestimated". This shows the importance of adopting the robust and probabilistic approaches.

Then, we compare the results of the following three columns, which represent the three cases employing $B(= 100)$ compatible classification models. For the simple case results, there are some ameliorations in the group, whereas there is one NPP (x_{16}) that is assigned to a worse category than before. This is explained as follows. In the procedure of majority voting, the

Table 8: Resume of after-action assignments of the considered NPPs with budget constraint and $B = 100$. White cases of the third to the sixth comlumns indicate unchanged assignments.

Budget constraint: Bg=40	original assignments by single "optimum" model	original assignment by MV	simple optimization assignment by MV	Robust optimization assignment by MV	Probabilistic optimization assignment by MV
x1	3	3	3	3	3
x2	2	3	3	2	2
x3	3	3	2	2	2
x4	3	3	3	3	3
x5	2	3	3	3	2
x6	2	2	2	2	2
x7	4	4	4	4	4
x8	4	4	4	4	3
x9	4	4	4	4	4
x10	2	2	2	2	2
x11	3	3	3	3	3
x12	3	3	3	3	2
x13	4	4	3	3	3
x14	3	3	3	3	3
x15	2	4	4	4	3
x16	3	3	4	3	3
x17	3	4	4	4	3
x18	2	3	3	2	3

number of models that originally assign x_{16} to A^3 is slightly lower than that to A^4. With the actions obtained by the simple optimization, some models that originally assigned x_{16} to A^3, now evaluate it in category A^2; at the same time, the number of models that assign it to A^4 does not change, becoming the majority. This further calls for the adoption of robust and probabilistic approaches. Indeed, the robust case gives a better result than the simple one. No NPPs are assigned to a worse category as before (x_{16}). More NPPs are improved (x_2 *and* x_{18}) but there is still only one NPP that is ameliorated from the worst category (A^4).

Finally, the probabilistic case shows a more promising way to choose the set of protective actions. Actually, 8 out of 18 NPPs are ameliorated. Among these, 4 were originally assigned to A^4, whereas the other 4 were originally assigned to A^3. This matches well our "expected use" of the limited resources ($B_g = 40$), implicitly defined by the weighted objective function I^x.

5.2 *Explanation of the behavior of the robust and probabilistic optimizations*

As presented in the previous subsection, the robust case shows a better amelioration performance than the simple one. However, the robust approach is in principle expected to ameliorate the group performance by giving preferential consideration to the NPPs that are originally assigned to the worst category: in this view, it does not provide a satisfactory performance, especially because it does not improve as many NPPs in category 4 as the probabilistic approach

does.

For further understanding, additional calculations and verifications have been carried out. Based on the original assignments obtained by majority voting (results in column 2 of Table 8) and considering $B(=100)$ compatible classification models, we choose all the NPPs that are assigned to the worst category (A^4) as a reduced study set $x^{worst} = \{x_p, p \in \{7, 8, 9, 13, 15, 17\}\}$. Then, imposing a limited budget ($B_g = 40$, as in the previous calculations) and an unlimited budget (conveniently set in the numerical algorithms to $B_g = 1000 >> B_{g_{min}} = 78$), we apply the robust and probabilistic optimizations to ameliorate the set x^{worst}. It turns out that for both cases, no feasible solution can be found by the robust algorithm. It means that considering only the NPPs originally assigned to the worst category (A^4), there is no way to generate a set of actions that can produce an amelioration of the group of plants for any of the $B(=100)$ compatible models. In other words, for any set of actions and considering all the $B(=100)$ compatible models, there is at least one after-action objective function value equal to "0".

In order to verify this hypothesis and find out for how many of the bootstrapped models the objective function value is "0", we use the optimal set of actions obtained in the robust optimization for the full set of alternatives (x) and the same budget limit ($B_g = 40 \; and \; 1000$) to ameliorate the NPPs of x^{worst}, taking into account the $B(=100)$ classification models. A bootstrapped distribution of the objective function $I^{x^{worst}}$ is obtained (Figure 6, left). We repeat the same process with the optimal set of actions obtained for the probabilistic case (Figure 6, right). In order to have a clearer view of the results, the Figures show the values of the objective function for the bootstrapped models rearranged in increasing order. For the two cases and the two different budgets (limited and unlimited), the corresponding objective function values for some compatible models are always "0" (especially for the robust case with a limited budget, $B_g = 40$). In comparing the results based on the given budgets, although for some compatible models with the given set of actions the performance cannot be ameliorated, the results of the case of an unlimited budget ($B_g = 1000$) are better than the ones of a limited budget ($B_g = 40$). With the same budget, the results of the probabilistic case are always better than the robust case. Except for the robust case with limited budget ($B_g = 40$), for the other three tests the number of models for which no plant is ameliorated is 6. In addition, they are the same 6 models for the three cases (named M^{na}). It is due to their characteristics of lower profiles and to the weights that the NPPs in x^{worst} can never be elevated to a better category after any set of actions.

If we consider the full alternatives set (x), the two approaches always produce feasible solu-

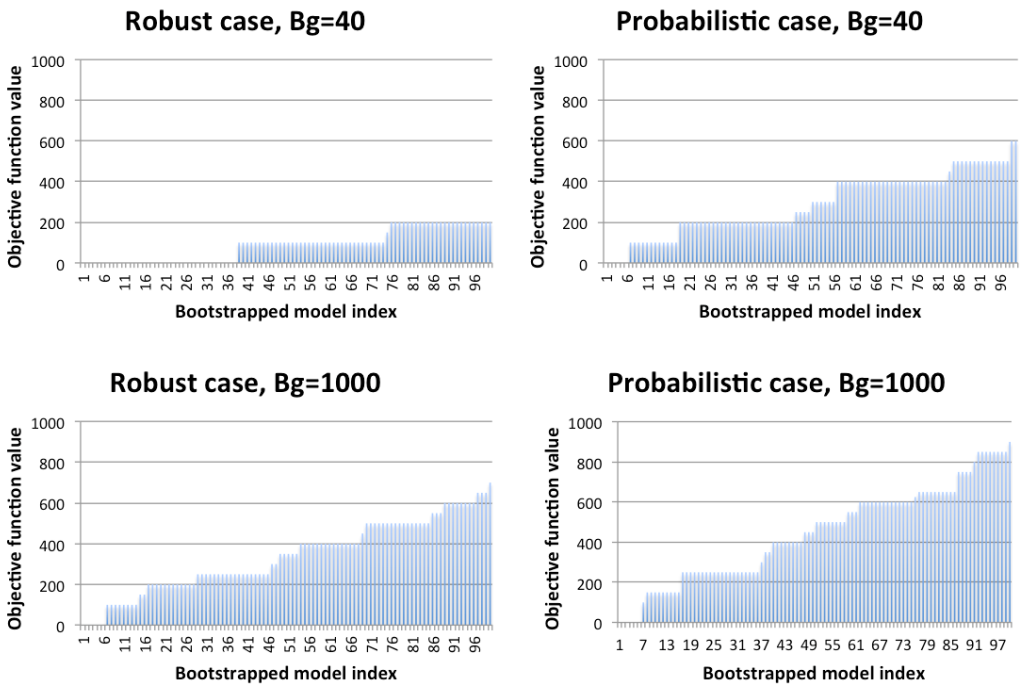

Figure 6: Distribution of after-action objective function values for x^{worst} in considering $B = 100$ classification models

tions. With the same budgets, similar calculations are done and the distribution of the objective function I^x of the whole group of NPPs can be obtained (Figure 7).

From Figure 7, we can find out that, except for the probabilistic case with $B_g = 40$, all the after-action objective function values are positive. Especially for the robust cases, if we go into the details of the ameliorations of the group, we can find out that, in correspondence of the 6 models in M^{na} that do not allow any after-action ameliorations for the NPPs of x^{worst} (mentioned in the previous paragraph), the after-action ameliorations for the whole NPPs group concentrate on the changes of the NPPs that are originally assigned to A^3.

Thanks to these new tests, we have a further understanding of the robust case and probabilistic case. The aim of our optimization is to find out a set of protective actions in order to ameliorate the group performance in giving a preferential consideration of the NPPs that were originally assigned to a relatively worse category. For the robust case, the amelioration of the worst case over all the models is demanded. In other words, the obtained set of actions should have a positive effect on the objective function value for any classification models. In this case, the optimization algorithm tries to improve the objective functions produced by those compatible models that produce the worst results and possibly presents particularly "pathological" features (in our case the models in M^{na}). Since in these configurations it is not possible to ameliorate the NPPs that are originally assigned to A^4, but we still force the algorithm to finally improve

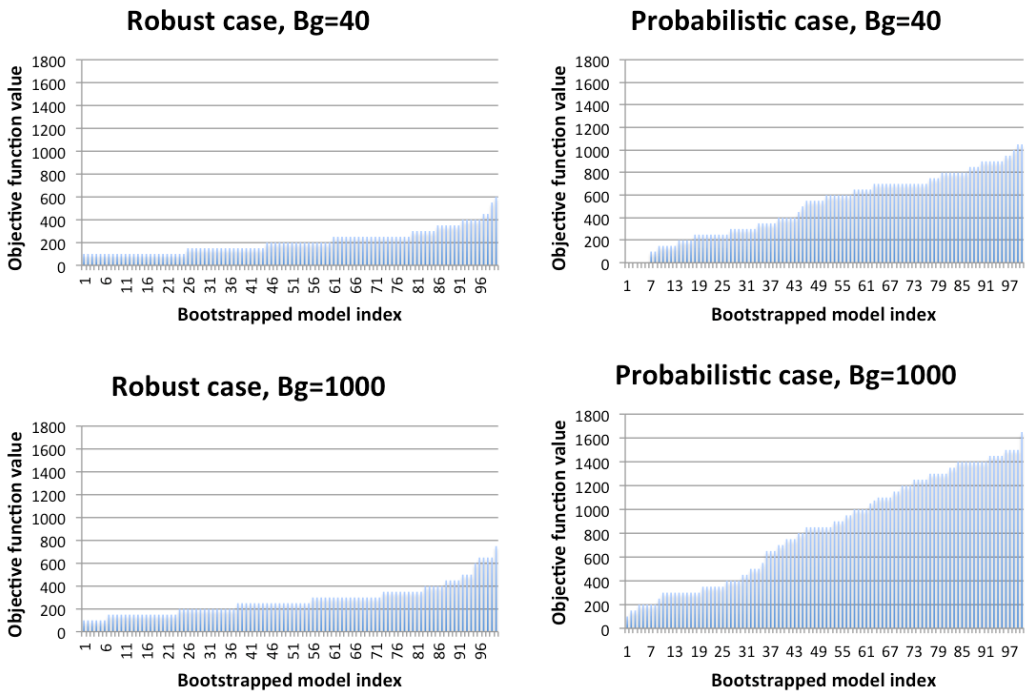

Figure 7: Distribution of after-action objective function values for x in considering $B = 100$ classification models

the objective function I^x, the algorithm then tries to target the NPPs that are originally assigned to A^3 (which is the second worst original category): actually, according to the definition of the objective function, increasing preference is given to NPPs with high vulnerability. This is also the reason why, if we apply this set of actions on the whole group of NPPs and evaluate them by the B models, the ones that are originally assigned to the worst category (A^4) are not improved as we might expect. On the contrary, for the probabilistic case, an expected value over B models is maximized. Since the amelioration of NPPs that are originally assigned to the worst categories is given a higher weight, the NPPs that are originally assigned to A^4 are more targeted. If the budget is not sufficient to improve at least one NPP for all models, such "extreme" models are "in practice" neglected by the algorithm. As a result, as presented in Figure 7, for the probabilistic case with a limited small budget, the expected value of the objective function is increased, but for some models, the corresponding value of function I^x is still "0".

For verification purposes, we considered a study of a new group of compatible models. We take the previous $B(= 100)$ compatible classification models and remove the models in the group M^{na} to form group M^a where the remaining $94(= 100 - 6)$ compatible models are considered. We repeat the optimizations with a budget limitation of $B_g = 40$ as before (subsection 5.1), with only the 94 models in M^a. The results are listed in Table 9. Comparing the results with those of (Table 8), it can be seen that in the first two columns they are exactly the same: for example,

Table 9: Resume of after-action assignments of the considered NPPs with budget constraint and $B = 94$. White cases of the third to the sixth comlumns indicate unchanged assignment.

	94 models				
Budget constraint: Bg=40	original assignments by single "optimum" model	original assignment by MV	simple optimization assignment by MV	Robust optimization assignment by MV	Probabilistic optimization assignment by MV
x1	3	3	3	3	3
x2	2	3	2 or 3	2	2
x3	3	3	2	2	2
x4	3	3	3	3	3
x5	2	3	3	3	2
x6	2	2	2	2	2
x7	4	4	4	4	4
x8	4	4	4	3	3
x9	4	4	4	4	4
x10	2	2	2	2	2
x11	3	3	3	3	3
x12	3	3	3	3	2
x13	4	4	3	3	3
x14	3	3	3	3	3
x15	2	4	4	4	3
x16	3	3	4	3	3
x17	3	4	4	3	3
x18	2	3	3	3	3

in the first column, we always consider the same single "optimum" classification model; also since there are only 6 models that are removed, the original assignments of all the NPPs by majority voting rule are not changed either. The results of the following columns show the same tendency as before (subsection 5.1).

There is one difference with the results of the simple case: x_2 is originally assigned to A^3 and in the bootstrap cases with $B = 100$, the after-action assignment is also A^3. However, with $B = 94$, the after-action assignment is either to A^2 or A^3, because the number of models that assign it to the two different categories remains the same. It means that the set of action generated by the simple optimization in considering the "optimum" classification model does not have a satisfactory influence under the evaluation of the models in M^{na} (assigned to A^3). The "optimum" classification model is generated based on the maximum number of pre-assigned training alternatives. The set of actions generated from this model is not robust enough but should be more efficient than a random one. The fact that the results based on the models in M^{na} are relatively worse proves the fact that these models are not as similar to the "optimum" one as the others. This is also the reason why the results of the robust case with $B = 94$ are better than the ones with $B = 100$. Firstly, one more NPP is improved (x_8, from A^4 to A^3) and there is one change of amelioration target: x_{17} is ameliorated (from A^4 to A^3) instead of x_{18} (from A^3 to A^2). This means that, after having removed the models in M^{na} from the considered group of compatible

classification models, the NPPs that are originally assigned to the worst category (A^4) are more likely to be improved based on the other 94 models. However, the overall results are still not as good as the probabilistic case (which remain the same as before). Although after the deletion of the models in M^{na}, the optimization algorithm will consider some other models that are less "extreme" than the ones in M^{na}, they still present a variability that may prevent the improvement of some NPPs to a better category. On the contrary, the probabilistic approach always searches to maximize the average value over the compatible models: thus, in practice, the models that give a poor output (i.e., the "tails" of the distribution) are given less importance.

5.3 *Rationality of the weighted-sum objective function in the probabilistic optimization*

As presented in the previous sections, the set of protective actions generated from the probabilistic case has a more satisfactory result than the others. A verification study is done with respect to this.

We consider the whole group of NPPs and the $B = 100$ compatible classification models. The weights used to represent the importance of changes from A^4 to A^3, A^3 to A^2 and A^2 to A^1 are $100, 50, 25$. Now, we change to a uniform set of weights $(100, 100, 100)$ and count the cumulative number of changes from each category to its adjacent improved category for all NPPs and all models. The changes are compared for each NPP between its original and the corresponding after-action category evaluated independently by each compatible model. A change of categories more than one for a NPP is considered as the combination of single changes from different "original" assigned categories (e.g., for one NPP that is ameliorated from A^4 to A^3, it is counted as one change from A^4 to A^3, one from A^3 to A^2 and one from A^2 to A^1).

We discover that, reasonably, with a very small amount of budget (e.g., $B_g < 10$), the algorithm can be infeasible since the given resource is not enough to make any amelioration. With a very large budget (e.g., $B_g > 702 = total\ cost\ of\ all\ highest\ level\ actions\ applied\ on\ all\ NPPs$), since the resource is adequate, there is no need for the weights to steer its allocation: for the different weights set, the final ameliorations are the same. Focusing on realistic cases of limited and relatively small budgets, we consider 4 different budgets, $B_g = 20, 40, 50\ and\ 90$ and run two optimizations with the weights of the objective function set at $(100, 50, 25)$ and $(100, 100, 100)$, respectively. The results are shown in Figure 8. It is obvious that, the bigger the budget, the bigger the number of cumulative changes after actions. For budgets $B_g = 40, 50\ and\ 90$, the

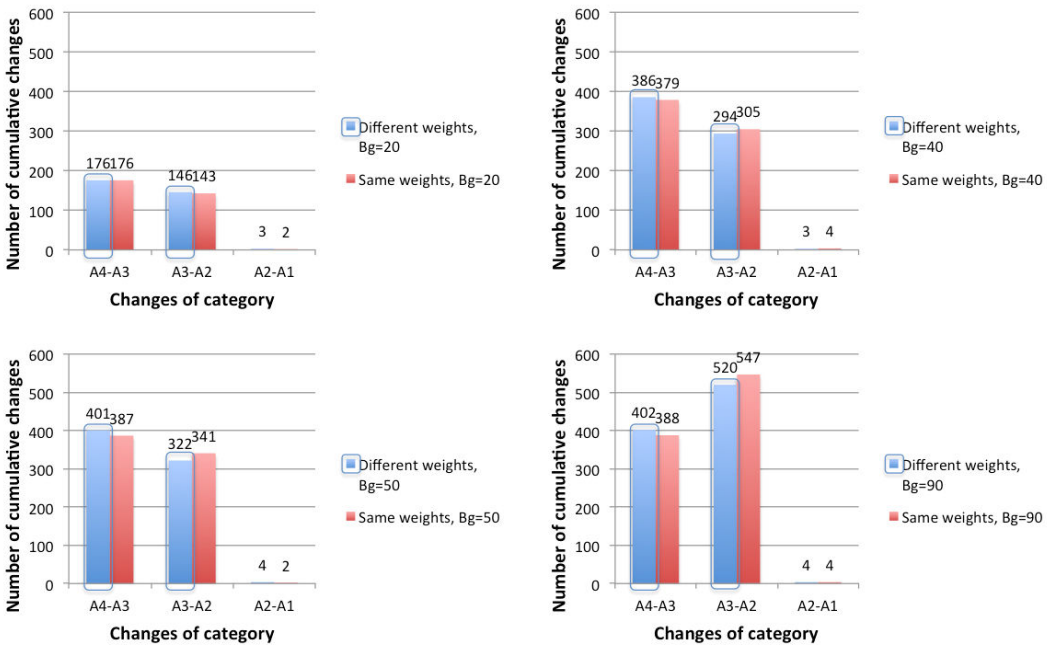

Figure 8: Distribution of number of cumulative changes of category for x in considering $B = 100$ classification models, two sets of weights of objective function and 4 different budgets

number of ameliorated NPPs originally assigned to the worst category (A^4) is bigger in the case of the objective function with weights $(100, 50, 25)$ than in the case of uniform weights $(100, 100, 100)$. For the budget $B_g = 20$, the number of ameliorated NPPs originally assigned to the worst category (A^4) is the same with both weights sets. But with the set $(100, 50, 25)$, a larger number of NPPs originally assigned to the second worst category (A^3) is ameliorated than with uniform weights $(100, 100, 100)$.

These results show that the probabilistic optimization with the objective function that contains a set of priority weights $(100, 50, 25)$ is, indeed, able to steer the allocation of the linked budget on protective actions that can ameliorate the performance of the NPPs, with preferential consideration to those originally assigned to worse categories.

6 CONCLUSIONS

We have addressed the issue of selecting a set of protective actions for minimizing the vulnerability of safety-critical systems (in the case study, nuclear power plants), within an optimization framework based on an empirical classification model. In particular, an MR-Sort model trained by means of a small-sized set of data representing a priori-known classification examples has been used.

Three optimization approaches have been developed and investigated: (i) one single classifi-

cation model is built to evaluate and minimize system vulnerability; (ii) an ensemble of compatible classification models, generated by the bootstrap method, is employed to perform a "robust" optimization, taking as reference the "worst-case" scenario over the group of models; (iii) a distribution of classification models, still obtained by bootstrap, is considered to address vulnerability reduction in a "probabilistic" fashion (i.e., by minimizing the "expected" vulnerability of a fleet of systems). To the best of the authors' knowledge, it is the first time that an inverse classification problem is formulated and considered for the optimization of the choice of protective actions to reduce the vulnerability of a group of safety-critical systems (e.g., Nuclear Power Plants), taking into account the uncertainty associated to the classification models.

From the results obtained, it can be concluded that a combination of protective actions can be obtained using only a single classification model, but this set of actions is not robust with respect to the uncertainty of the classification model. The robust optimization may, then, be used for a more conservative set of actions, coping with model uncertainty. Eventually, the probabilistic optimization seems most practical for real cases, for the following reasons: (i) as for the robust case, it handles the uncertainty coming from the finite data set available and the compatible models; (ii) by maximizing the expected value of the bootstrapped probability distribution of the objective function, some " extreme" compatible models of the bootstrapped ensembles are "neglected", which is reasonable and more realistic.

REFERENCES

Aggarwal, C. C., C. Chen, & J. W. Han (2006). On the inverse classification problem and its applications. *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference. IEEE*, 111–113.

Aggarwal, C. C., C. Chen, & J. W. Han (2010). The inverse classification problem. *JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 25*, 458–468.

Aven, T. (2003). *Foundations of Risk Analysis*. N.J-Wesley.

Aven, T. (2010). Some reflections on uncertainty analysis and management. *Reliability Engineering & System Safety 95*, 195–201.

Aven, T. & R. Flage (2009). Use of decision criteria based on expected values to support decision-making in a production assurance and safety setting. *Reliability Engineering & System Safety 94*, 1491–1498.

Belton, V. & T. Stewart (2002). *Multiple Criteria Decision Analysis–An Integrated Approach*. Kluwer Academic Publishers.

Doumpos, M. & C. Zopounidis (2002). *Multicriteria Decision Aid Classification Methods*. Netherlands: Kluwer Academic Publishers.

Efron, B. & R. Thibshirani (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Kröger, W. & E. Zio (2011). *Vulnerable Systems*. UK: Springer.

Leroy, A., V. Mousseau, & M. Pirlot (2001). Learning the parameters of a multiple criteria sorting method based on a majority rule. *The Second International Conference on Algorithmic Decision Theory*, 219–233.

Li, A. G., x. Zhou, & J. L. Zhang (2012). Performance analysis of quantitative attributes inverse classification problem. *JOURNAL OF COMPUTERS 7*, 1067–1072.

Mousseau, V. & R. Slowinski (1998). Inferring an electre tri model from assignment examples. *Journal of Global Optimization 12*, 157–174.

Roy, B. (1991). The outranking approach and the foundations of electre methods. *Theory and Decision 31*, 49–73.

Wang, T. R., V. Mousseau, N. Pedroni, & E. Zio (2014). Assessing the performance of a classification-based vulnerability analysis model. *Risk Analysis doi:10.1111/risa. 12305*.

Wang, T. R., V. Mousseau, & E. Zio (2013). A hierarchical decision making framework for vulnerability analysis. *Proceedings of European Safety and RELiability Conference (ESREL 2013)*, 1–8.

Zio, E. (2006). A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. *IEEE Transactions on Nuclear Science 53*, 1460–1470.