MS-E2177
Seminar on Case Studies in Operations Research
Expert judgements for cost assessment in risk management
Final Report

Jani Laine
Jani Mäkinen (Project Manager)
Sampo Riekki
Lauri Sääskilahti

May 30, 2022

Contents

1 Introduction

1.1 Background

Our client Inclus is a Finnish technology start-up company founded in 2012. Inclus develops user-friendly, visual and interactive cloud software for risk management. Their methods consist of multicriteria and network analysis and their software is used world-wide in different fields such as risk management for large-scale construction projects, trend and competence analysis in public sector organizations and conflict prevention in international crisis management.

Risk management means coordinated activities to direct and control an organization with regard to risk [1]. It includes analyzing the risks and the means to deal with the risk appropriately. Risk analysis consists of a process of assessing the likelihood of an adverse event occurring as well as its potential impact. Risk management can be qualitative or quantitative, and often involves the use of mathematical models or expert opinions or both. Qualitative models give an explanation of the risk and relating factors whereas a quantitative model tries to quantify the risk and assign a numerical value to it. The construction of both models require information the accuracy of which significantly affects the model's performance.

1.2 Motivation

Information for risk analysis and management is often collected by eliciting it from experts. It is common to use two criteria for the elicitation, impact and probability. Impact of an event can be expressed by for example a Likert-based classification such as an integer value between 1 and 4 or by the financial cost the event causes. Probability specifies how likely the event is. The two criteria are often multiplied together for a given risk factor by some convention and the resulting value will correspond to the seriousness of the given risk factor. Having only these two values for the analysis of risk, however, is limiting in that they do not capture all aspects of risk.

Another common problem in eliciting information from experts is that they usually tend to give only one value per criterion. A single value per criterion implies that the impact and probability are known for certain. This is rarely the case since the information consists of predictions and estimates of the experts and thus have uncertainty in them. Failing to take uncertainty into account can be detrimental to the model's performance.

These limitations of the information make it hard to create simulations of risk and estimations of the total risk of a project. Also, the absence of knowledge of any correlation or interrelations between risks further hinders simulations. Clients in the field of risk management need more advanced and robust tools for managing risk without losing user-friendliness in the process. Our project explores ways to incorporate uncertainty and interdependencies of risk into the risk managing model within the Inclus context.

1.3 Objectives

Our aim is to give sound and scientifically founded suggestions for improving the elicitation process used by Inclus. The objective is to conceptualize our suggestions, which may later be implemented by Inclus. Developing working code to achieve the functions included in the suggestions is not a part of the objective.

Our project can be divided into two distinct parts: in the first part we develop an approach on how to best capture the uncertainty related to the risks. Single-value probability-impact estimates have their flaws as they leave out interesting and useful information. Thus we demonstrate and adapt the well-known three-point estimation procedure from the scientific literature to elicit more information about the distributions. We also try and experiment more a visual alternative to the three-point estimates.

In the second part our aim is to conceptualize an approach on how to elicit information about the interdependencies of different risks. All the risks involved are rarely independent, and treating them as such makes some amount of oversimplification, reducing the accuracy and correctness of the analysis. An example is a personal injury on a construction site, which naturally increases the probability of a sub-project to be delayed, which in turn could delay the whole construction project. For this purpose we conceptualize an approach reminiscent of a mind-map with a graph structure.

Besides eliciting information from individual experts we suggest an approach to effectively aggregate various expert judgements. This process is essential for both distinct parts of the project, as Inclus is almost always working with multiple-expert teams, and thus means for aggregating judgements is necessary.

Our objectives also take into account the specific needs of Inclus. For example the code of the current software version needs to run on the end user's web browser, limiting the complexity of the calculations. Also Inclus has focused more on simpler, easier-to-interpret methods rather than heavy Bayesian Network approaches more typical to safety critical systems. Thus our aim is to complete the objectives so that Inclus can easily incorporate them into their existing workflow.

As a whole the this report will yield a systematic approach to eliciting uncertainties about the studied system in a meaningful way that acknowledges the interrelations of different risks.

2 Literature review

2.1 Eliciting and aggregating expert opinions

The results of both probability estimation and financial cost estimation presented in this project are solely based on expert judgements. A proper protocol for eliciting expert judgements needs to be presented in order to obtain reasonable outcomes. One method of improving elicitation and aggregation quality is a structured expert judgement elicitation procedure called IDEA protocol presented in article by Henning et al. [2]. The IDEA acronym stands for the following key steps of the protocol: Investigate, discuss, estimate and aggregate. It is an iterative elicitation process where experts are asked anonymously the same questions twice (or more if necessary). The first round is for standardization of the question parameters such that experts agree on the magnitudes and the meaning of the questions. The visual showcase of these results is followed by the second elicitation round, where experts now have the common base for answering the questions properly. The phases of the IDEA protocol are visualized in Figure 1.

Preparing for the IDEA protocol includes recruiting a diverse group of experts that are assumed to have considerable knowledge to the subject. The process of selecting the most suitable experts for elicitation is far from trivial. A common pitfall is to assess experts based on external factors alone, such as age, publications and peer recommendations. Much better option is to gather empirical evidence of expert performance on closely related tasks and queries. This information is not often available, but it should be used as a priori whenever possible. The main criterion is that experts understand the questions being asked. It is also important to have a sufficient diversity to the group of experts, that should reflect variation in factors such as gender, age, education and cultural background. Next step is to decide the elicitation format, which in the purposes of this project will be remote elicitation. It can be done in the form of online query where experts are asked for probabilistic or quantitative responses, for which the protocol incorporates two alternative question formats depending on what is elicited. These formats are three-step elicitation and four-step elicitation corresponding Likert-based probability estimation and impact estimation for financial cost of risks respectively. Remote elicitation is also the best and simplest option for preserving expert anonymity, which is important for the steps of IDEA protocol. The documents need to have clearly stated questions with unique interpretations such that different experts agree on the meaning of the questions and units. [2]

The main phase of the IDEA protocol begins with the investigation, where all experts answer the questions anonymously, individually and without communicating with each other. Their responses are values on a continuous scales with an option to provide verbal reasoning for their judgements. Lowest, highest, most likely values and probabilities, and confidence intervals for quantities are given. The confidence interval represents the estimated probability that the quantity lies between the lowest and highest values and it is somewhat reflecting to expert's certainty of the estimate. After the first elicitation round, the data might require some necessary cleaning, such as removing incorrect answers where lowest values are higher than the highest values. [2]

Next the intervals are standardized if the level of credible intervals were given. Typical credible interval values are 80 or 90 percent. Standardization is calculated with the following linear extrapolation,

$$I_L = B - \frac{S}{C}(B - L)$$
$$I_H = B + \frac{S}{C}(U - B) \tag{1}$$

where B is the most likely estimate, L is the lowest estimate, U is the highest estimate, C is the confidence interval given by expert and S is the level of credible intervals to be standardized to. [3] To finally calculate the group aggregate estimate, quantile aggregation is performed by calculating arithmetic means of the lowest, highest and most likely estimates. However, the performance of quantile aggregation is no as strong as aggregating the data to fitted distributions, which is shown in the study by Colson and Cooke [4]. Fitting to distributions are further investigated in the next chapter.

The results of the first round are graphically displayed to the experts, which leads to the discussion phase. Participants are shown the group aggregate estimates of the outputs as well as the individual answers labeled with their code names. Discussion could explore for example the sources of variation in the answers. This phase is more prominent in the case of non-remote elicitation, since it allows free communication between participants.

The same questions are asked in round 2, where the participants have the graphical outputs of the first round as a priori. Experts can adjust their answers to correspond equal confidence intervals and outlying answers are more likely to shift towards the group aggregate estimate, unless the experts have a very strong confidence for their reasoning. Second round results are processed similarly to the first round, and corresponding graphical output is produced as the result. [2]

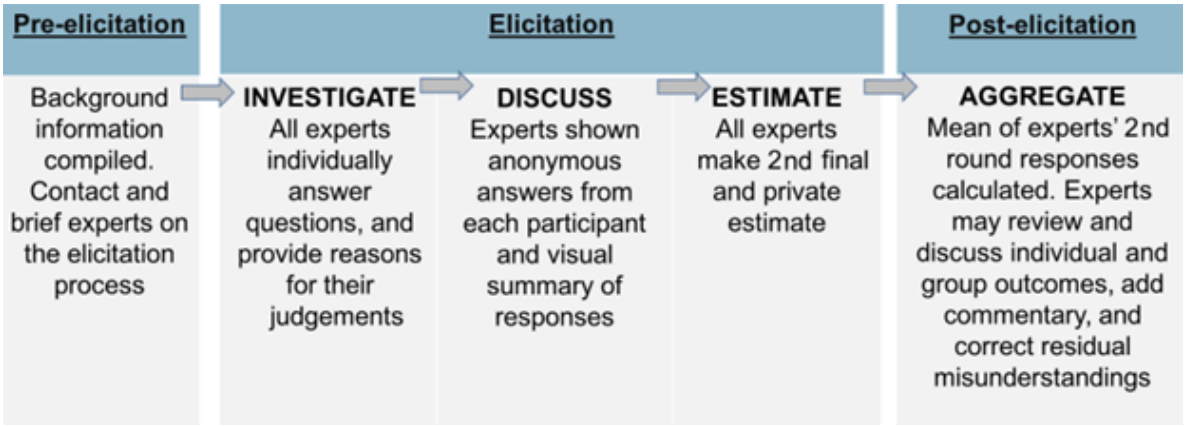

Figure 1: The phases of the IDEA protocol

2.2 Three-point estimates

The elicitation of three-point estimates, the expert is requested to provide best-case estimate, the most-likely estimate, and the worst-case estimate of the impact of a given risk. The motivation is that these estimates can be used to construct a probability distribution for a given impact. A single-value estimate implies that the impact of a given risk is known, which is often not the case. Furthermore, single-value estimates cannot express how the probability of an impact is distributed. To illustrate, consider the two following cases. In the first case the most-likely estimate and the best-case estimate are relatively close to each other, but the worst-case estimate is relatively far. In the second case the most-likely estimate and the worst-case estimate are relatively close to each other, but the best-case estimate is relatively far. Also, assume that in both cases the most-likely estimates are the same. Should an expert answer either question with a single-point estimate, we should expect to find both estimates very close to each other, even though the impacts behave differently. When asked for a single-point estimate, the expert is not primed to think about how the impact is distributed. The expert just tries to answer what the impact is given that the risk becomes reality.

A three-point estimate has the ability to tell a richer story by taking in more information about the impact and turning it to a probability distribution. However, it should be pointed out that three-point estimates are not silver bullets for risk assessment. The level of accuracy depends on two error prone factors: the accuracy of the elicited points, and the correspondence of the assumed distribution with the reality. We also need to appreciate the fact that eliciting three-points requires more attention and deliberation from the expert, leading to a more laborious questionnaire.

The literature review revealed three distributions that can be utilized in three-point estimates: triangular distribution, PERT distribution, and the generalized two-sided power distribution. These distributions are often found in risk analysis as a way to quantify uncertainty of task duration [5]. The next chapters briefly introduce these distributions.

For simplicity, the range of possible values for an impact is $[0, 1]$ in the following examples. However, any arbitrary range can be used with any of the following distributions simply by normalizing the data between $[0, 1]$, fitting the distribution to the data, and then then converting it back to the original scale.

2.2.1 Triangular distribution

Triangular distribution is defined by the probability density function,

$$p(x|a,b,c) = \begin{cases} 0, & \text{for } x < a, \\ \frac{2(x-a)}{(b-a)(c-a)}, & \text{for } a \leq x \leq c, \\ \frac{2(x-a)}{(b-a)(c-a)}, & \text{for } c < x \leq b, \\ 0, & \text{for } x > b, \end{cases} \tag{2}$$

where a corresponds to the minimum value, c to the mode and b to the maximum value of the distribution. In a three-point estimate, these values would correspond directly to the values elicited from an expert. As a consequence, triangular distributions are very easy to create from data. The probability density function and cumulative distribution function of a triangular distribution can be seen in Figure 2 with the minimum, mode and maximum values indicated by the red dashed lines.

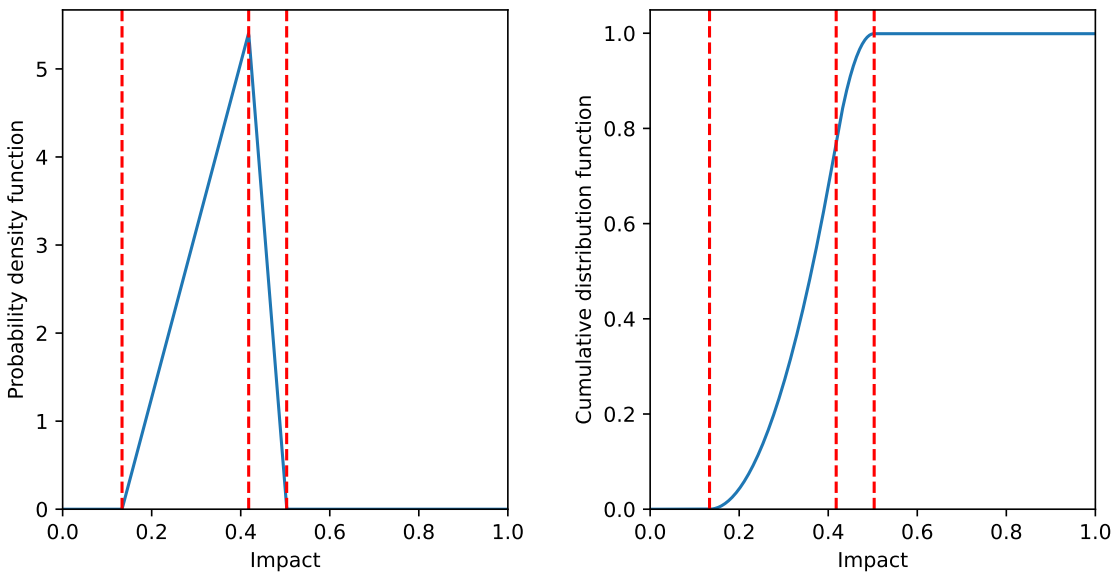

Figure 2: The probability density and cumulative distribution functions of a triangular distribution.

2.2.2 PERT distribution

The name of the PERT distribution is abbreviated from "Program Evaluation and Review Technique", which is a project management tool used to analyze how tasks of a project will be completed over

time. The tool was first developed by the U.S. Navy Special Projects Office for a nuclear submarine project [6]. PERT distribution is defined by the probability density function,

$$p(x|a,b,\alpha,\beta) = \begin{cases} 0, \text{ for } x < a, \\ \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta+1}}, \text{ for } a \leq x \leq b, \\ 0, \text{ for } x > b, \end{cases} \tag{3}$$

$$\alpha = 1 + 4\frac{c-a}{b-a}, \tag{4}$$

$$\beta = 1 + 4\frac{b-c}{b-a}, \tag{5}$$

where B is the Beta function, a corresponds to the minimum value, c to the mode and b to the maximum value of the distribution. The PERT distribution is a transformation of the Beta distribution with the assumptions that the distribution is defined between the continuous interval of $[a, b]$ and that the expected value of the distribution equals $\mu = (a + 4c + b)/6$. The mode is weighted four times as much as the end-points, however, this choice is more of a convention than a necessity. This convention stems from the suggestion of the original authors that the standard deviation of the distribution should be $1/6$ of the support $[a, c]$. [6] The default PERT distribution places more weight on the mode parameter, compared to the triangular distribution that has an expected value of $\mu = (a+b+c)/3$. The probability density function and cumulative distribution function of a PERT distribution can be seen in Figure 3 with the minimum, mode and maximum values indicated by the red dashed lines.

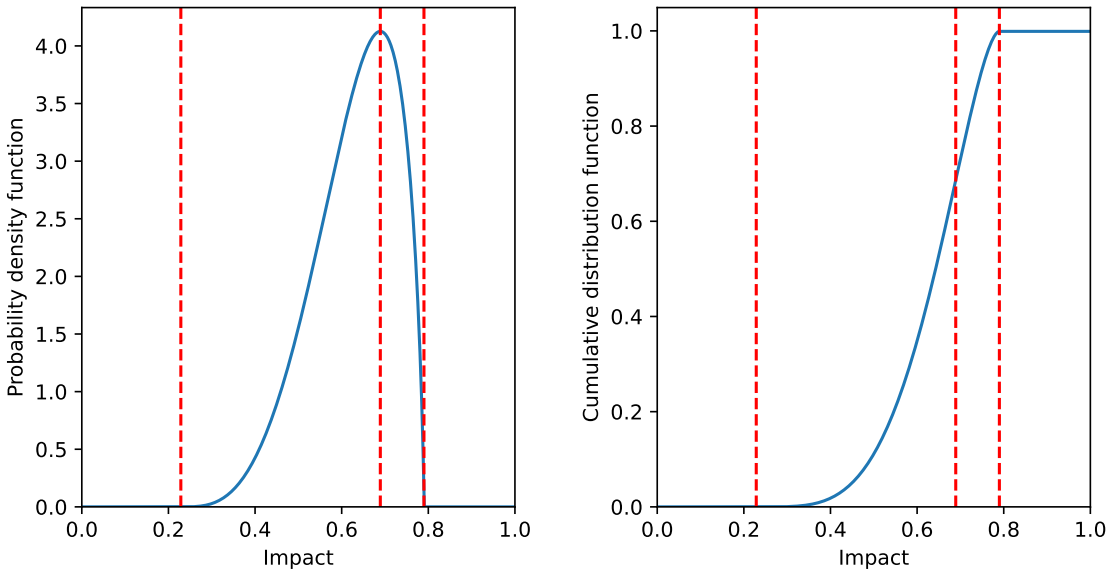

Figure 3: The probability density and cumulative distribution functions of a PERT distribution with $\gamma = 4$.

A modified PERT distribution uses an additional parameter γ to define the weight given to the mode in the expected value $\mu = (a + \gamma c + b)/(\gamma + 2)$ [7], leading to the modified PERT distribution, that has a probability density function of the form,

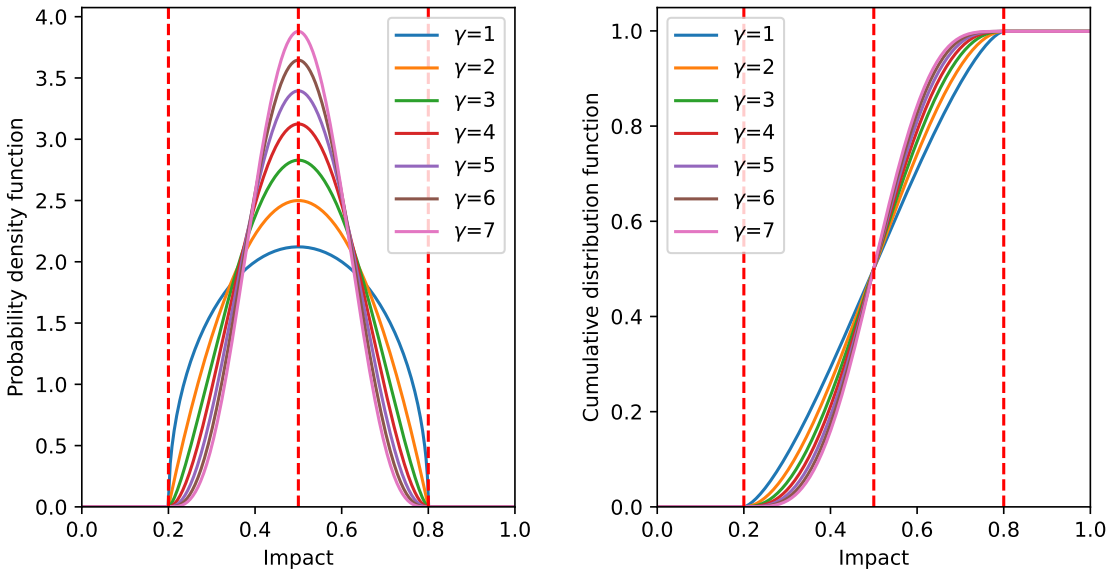

Figure 4: The probability density and cumulative distribution functions of a PERT distribution with γ ranging from 1 to 7.

$$p(x|a,b,\alpha,\beta) = \begin{cases} 0, \text{ for } x < a, \\ \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta+1}}, \text{ for } a \leq x \leq b, \\ 0, \text{ for } x > b, \end{cases} \tag{6}$$

$$\alpha = 1 + \gamma\frac{c-a}{b-a}, \tag{7}$$

$$\beta = 1 + \gamma\frac{b-c}{b-a}. \tag{8}$$

This extension allows the user to control how strongly the distribution should be centered around the mode and how much probability should be assigned to the tails of the distribution. The effect of parameter γ is shown in Figure 4, where its value was varied from 1 to 7. We can see from the Figure, that when $\gamma = 1$ the distribution is very tail heavy compared to when $\gamma = 7$. If the default value of $\gamma = 4$ is not used, typically values between $[2, 3.5]$ are used. The flattening effect on the distribution is especially useful for distributions where the relative distances to the mode from the extreme points is differs greatly. In Figure 5 we can see this phenomenon in action. With the default $\gamma = 4$ the distribution gives very low probability values for the long tail, whereas smaller γ values give the tail a noticeable increase in probability mass.

2.2.3 Generalized two-sided power distribution

The generalized two-sided power (GTSP) distribution is defined by the probability distribution [8],

$$p(y|m,n,\theta) = \frac{mn}{(1-\theta)m + \theta n} \times \begin{cases} \left(\frac{y}{\theta}\right)^{m-1}, & \text{for } 0 < y < \theta, \\ \left(\frac{1-y}{1-\theta}\right)^{n-1}, & \text{for } \theta \leq y < 1, \end{cases} \tag{9}$$

where $0 < \theta = c < 1$ is the mode of the distribution and $m, n > 0$ are power parameters that define the shape of the left and right sides of the distribution, respectively. Note that with $n = m = 1$ the GTSP distribution reduces to the uniform distribution, with $n = m = 2$ the GTSP distribution reduces to the triangular distribution, and with $n = m$ the GTSP distribution reduces to the two-sided power distribution. The GTSP distribution was suggested by Herrerías-Velasco et al. as a more flexible alternative to the four-parameter beta distribution [9].

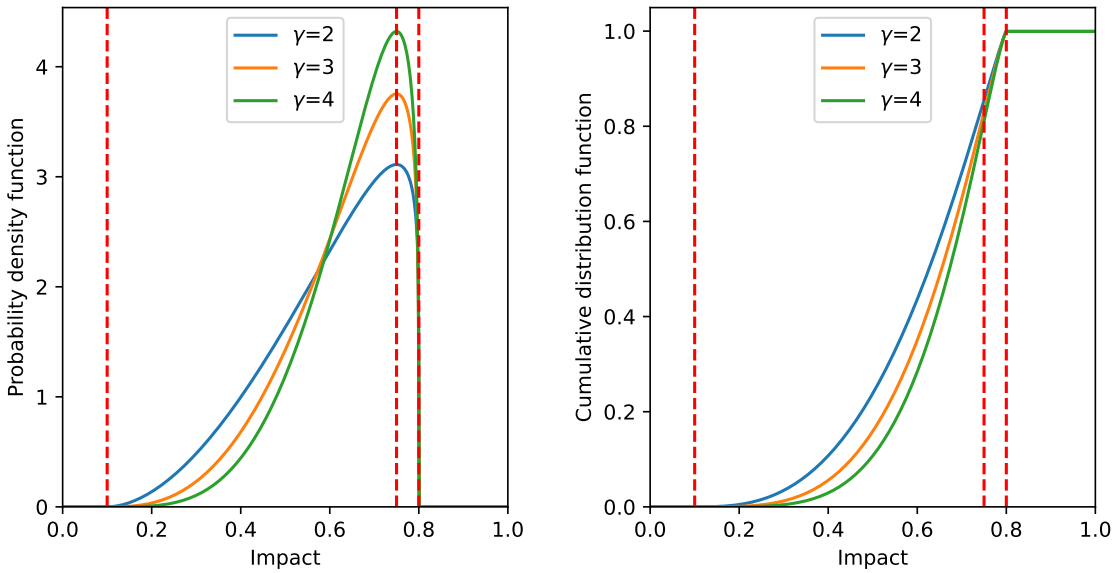

Figure 5: The probability density and cumulative distribution functions of a skewed PERT distribution with γ ranging from 2 to 4.

The probability density function and cumulative distribution function of a GTSP distribution can be seen in Figure 6 with the minimum, mode and maximum values indicated by the red dashed lines.

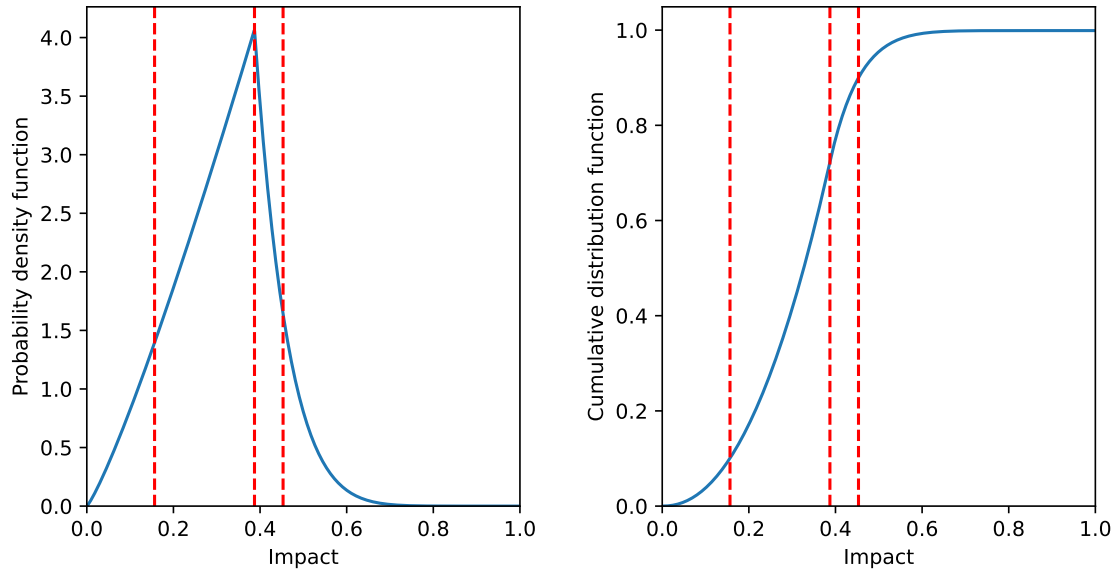

Figure 6: The probability density and cumulative distribution functions of a generalized two-sided power distribution with quantiles $p = 0.10$ and $r = 0.90$.

To solve the power parameters n, m from an elicited opinion a, b, c, we first need to define what quantiles these estimates correspond to. Contrary to the two previous distributions that had a probability mass of 0 outside the end-point estimates, the GTSP distribution has a non-zero probability mass on the whole support $[0, 1]$. For example, Herrerías-Velasco et al. propose quantiles $p = 0.10$ and $r = 0.90$, although they mention that other popular values, e.g. $p = 0.05$ and $r = 0.95$ will work equally well [9]. To solve the power parameters from an elicited opinion and selected quantiles, the

authors propose Algorithm 1. In practice, the algorithm solves the cumulative distribution function repeatedly for the lower and upper quantiles by keeping the power parameter responsible for the value of the given quantile as a variable and the other power parameter constant. The power parameters are solved repeatedly until their values converge close enough to the true values.

Algorithm 1 An algorithm for solving the power parameters of a GTSP distribution

1: $n^* \leftarrow 1$
2: Solve m^* from $F(x_p|\theta, n^*, m) = p$
3: Solve n^* from $F(x_p|\theta, n, m^*) = r$
4: **if** $|F(x_p|\theta, n^*, m) - p| < \epsilon$ **then**
5: **stop**
6: **else**
7: **go to** 2
8: **end if**

To illustrate how choosing the quantile effects the shape of the distribution, in Figure 7, we varied the quantile values between $[0.0125, 0.1]$. Symmetric ($p = q, r = 1 - q$) quantiles were used in all cases.

As a side-note, because the quantiles can be chosen freely, using GTSP distribution opens up the possibility of asking for quantiles instead of minimum and maximum values. However, in most cases we assume that this is not recommended and would lead to more confusion rather than more accurate estimates. It is not obvious that even people with mathematical backgrounds would provide more precise estimates by asking for quantiles. When the minimum and maximum values are being asked, we suggest using quantiles $p = 0.025$ and $r = 0.975$ to ensure the tails do not contain too much of the total probability mass.

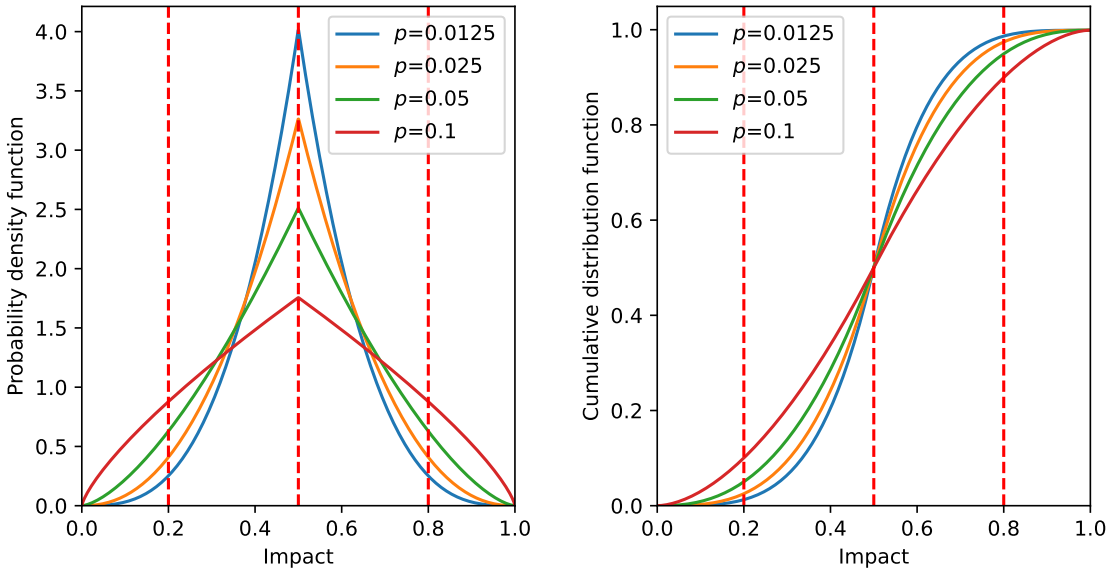

Figure 7: The probability density and cumulative distribution functions of a generalized two-sided power distribution with varying quantiles.

2.3 Risk dependencies

In statistics, when two events are probabilistically dependent, knowledge of the realization of one event affects the probability distribution of the other. There are many ways to describe the dependency between two events e.g. correlation, causation and cross-impact. The terms of probabilistic dependencies such as correlation are often used interchangeably with each other, and they may be confusing at times. Correlation in this text and in most scientific papers refers to a linear dependence and it is calculated as follows:

$$\rho(X,Y) = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\sigma^2[X]\sigma^2[Y]}}, \tag{10}$$

where $\mathrm{Cov}[X,Y]$ is the covariance between X and Y and $\sigma^2[X]$ and $\sigma^2[Y]$ the variance of X and Y, respectively. Correlation gets values in the interval $[-1,1]$. Negative values i.e. negative correlation means that when one of the variables increases the other one decreases. Positive values i.e. positive correlation means that when one of the variables increases the other one increases as well. The magnitude of the value represents the strength of the correlation and a value of zero means that there is no correlation between the values.

Causal relationship between two events means a dependency in which the occurrence of the first event affecting the occurrence of the second event. A causal relationship is only possible when the two events do not occur at the same time. For example, a statement such as "the event A may have caused the event B" would imply a causal relationship between the events and is according to [10] generally interpreted as meaning that the event A occurring increases the probability of the event B occurring i.e.

$$P(B|A) > P(B), \tag{11}$$

The cross-impact is another way of describing the dependency of two events by cross-impact analysis (CIA). It is designed for pairwise comparison of multiple events and to figuring out if and how they would impact future events. The method has been further improved in [10] to a probabilistic CIA (PCIA) in which the objective is to map a joint probability distribution across all possible scenarios of events of a system. The resulting probability distribution can then be used for the assessment of risk of the system. The cross-impact multipliers for the PCIA are defined as

$$C_{ij}^{kl} = \frac{P(A_i^k|A_j^l)}{P(A_i^k)}. \tag{12}$$

The multiplier describes how much more likely is the occurrence of B if the event A occurs with certainty. The information of the dependencies of events can be obtained from historical data or by elicitation from experts. Often historical data is not available or lacks in quality and thus elicitation is the only sensible way of obtaining the information [11]. The elicitation of dependency information is not a trivial procedure as it is based on opinions of people. Elicitation can have many caveats and the elicitation process should aim to mitigate them and gather the information as accurately as possible. There are more research papers that consider univariate information elicitation from experts in general such as [12] but very few that cover the elicitation of multivariate information i.e. dependencies [11]. In [11] the generalization of a univariate information elicitation process to fit the multivariate alternative is analyzed. The results of the analysis showed that the generalization is feasible in some ways but not wholly. In [13] the experts are encouraged to try the following thought experiment to prepare for the elicitation.

The expert is to estimate the likelihood of two variables A and B by ma and mb where ma and mb are medians of the assigned marginal distributions for A and B respectively. By definition, the expert should think that it is equally likely for ma to be lesser or greater than A i.e. underestimating the variable should be equally likely than overestimating it. The same applies for mb and B. Let's assume that the expert learnt that she or he had underestimated ma. Would this cause the expert to think that she or he had underestimated mb as well? If the expert now thinks that mb is not equally likely to be greater or lesser than B then it means that the expert judges A and B to be dependent.

As can be seen from the description of the thought process the elicitation process is prone to errors and biases. In conclusion [11] identifies two main difficulties with the elicitation process. First one is that the amount of information will burden the experts too much and the second one that the experts struggle to keep in mind and process complex scenarios. After the elicitation [11] proposed a scoring

system to assess the elicited information based on observation if available. The scoring would compare the elicited information to an existing distribution by e.g. calculating the Hellinger distance which is used for quantifying the similarity between two probability distributions.

After the expert opinions have been elicited and validated using some scoring method the newly gathered information must be used intelligently in order to benefit from it. The main objective of eliciting dependence information is to quantify a multivariate stochastic model when this cannot be done by other means [11]. There are different ways of modeling the risks one of them being to model the dependencies directly e.g. using Bayesian networks where the variables are represented by nodes in the graph and dependencies by arcs. An example of a simple Bayesian network is presented in Figure 8.

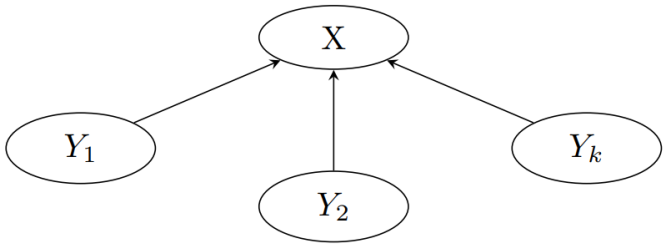

Figure 8: A simple Bayesian network with one child node X and three parent nodes Y_1, Y_2 and Y_k.

Often, there are more than one expert that the information is elicited from and thus aggregation is needed. The aggregation can be done by either behavioral or mathematical ways the latter of which is usually preferred in order to avoid errors that might be produced by social interaction [11]. The problem with mathematical ways is that some dependence information might be lost in the process. Two methods are generally used for mathematical aggregation: Bayesian aggregation and pooling methods. Bayesian aggregation methods provide better resilience for biases and pooling methods are usually easier to use [14].

2.4 User Interface design principles

The user interface (UI) for information input is used to allow the experts to input their opinions on the dependencies between the chosen risks. Their job is thus basically to point out dependencies between any two risks and then decide how strongly they think the two risks depend on each other. The risks have been previously chosen by some other method which is not included in this process. There exist some guidelines that should be followed when designing a UI for the information elicitation process.

The human mind has the so called visceral response to everything it perceives. This puts a lot of emphasis on the immediate perception of the UI [15]. The UI needs to communicate instantly what it is there for and how it should be approached. Therefore it should be obvious from the start where the information should be typed in or otherwise input and where the user can click to proceed [16]. Once the user decides to interact with the UI every action is done with some expectation which means that the UI should behave as intuitively as possible in order to avoid confusion.

In the process of the information elicitation, the users should be aware of their progress and where they currently are [16]. The sense of knowing the status of the process will make it easier for the users to concentrate on the task at hand. This can be achieved e.g. with a simple progress bar. Navigating should be easy and there should be an option to revisit previous pages as well as to undo previously made inputs [15].

Even though the elicitation process deals with experts who are used to giving judgment, the human mind usually tries to find the easiest and quickest way to complete its tasks. This might mean in some cases that the expert is mostly scanning the tasks, especially if there is a lot of content to judge. Therefore, the UI should support scanning meaning that important parts should be highlighted, the headings bigger than the text body and the texts should generally be kept short [16]. The philosophy guiding the design should be that there is human error, only poor design. Prior to releasing the product there should be extensive usability testing.

The user interface for presenting the output follows mostly similar guidelines to the ones stated previously but there are some specific things to focus on. In presenting something the visual hierarchy of the pages is important. Related things should be nested together and separated clearly from other information [16]. Needless visual noise should be eliminated to keep the user's focus on the task at hand.

3 Benchmarking

There exist other software solutions for risk management besides Inclus. These software include the category of popular statistical analysis software (R, SPSS, Matlab and Risk), smart questionnaire software ZEF [17], data interface driven Workiva [18], risk modeling sofware (Archer [19] and Vose [20]), risk workshop assessment tool Resolver Ballot [21] and strategic planning tool Thinking Portfolio [22]. In this section these software are benchmarked i.e. reviewed and compared to the one that Inclus provides. Benchmarking provides insight to the industry of risk analysis software and helps to avoid creating something that has already been developed.

The popular statistical analysis software such as R, SPSS, Matlab and Risk provide a large variety of tools for risk analysis. However, most of the software are not strictly designed for risk analysis and thus require manual work to be done in order to apply it to risk analysis. Consequently, they require a lot of domain expertise but provide also flexibility to skilled users. On the othe hand, Risk is a Microsoft Excel add-on designed for risk analysis. It features a capability for Monte Carlo simulations of an event which basically means simulating through all possible scenarios and determining their likelihood of happening and a capability for sensitivity analysis which ranks the risk factors by their influence.

The standard statistical analysis software can be used for risk analysis but they require more expertise than the software that Inclus provides. They also have to be installed on a computer whereas Inclus's software runs in a cloud. Although very capable when used skillfully these software cannot match Inclus's ease of use and user interface. In addition these software do not provide any capabilities for eliciting information which is a major part of risk analysis.

ZEF is software for eliciting information through smart questionnaires and a fairly close alternative to Inclus. ZEF's questionnaires are designed to be be user-friendly and to elicit information accurately but they are not specifically targeted towards risk analysis. They feature a 2D slider for assessing two components in each question, for example impact and probability of a risk factor or quality and meaning in a customer feedback survey. They use quite simple mathematics in the questionnaires such as calculating the average and normalizing the variables in the questionnaires in order to make them more accurate. ZEF also provides some tools for analysing and presenting the results from the questionnaires and it even features some artificial intelligence in the form of natural language processing for sentiment analysis.

ZEF is similar to Inclus but the focus of ZEF is on the questionnaires for a variety of industries whereas Inclus is made strictly for risk management. Both of them provide visual and user-friendly interfaces but Inclus has more variety in visualizing the results from the questionnaires. Both of them can run their software online and it can be accessed via a browser.

Workiva is a platform which is designed to enable seamless data exchange between different parts of a company e.g. departments or teams. They also provide solutions for enterprise risk management and their main thing is the aggregation of data from multiple sources. The risk management package includes e.g. some basic visualization of risk such as heat maps. Workiva represents a different kind of approach to risk management software than Inclus and it offers more of a complete service package whereas Inclus offers user-friendly tools.

The risk modeling software such as Archer and Vose represent a more hardcore risk management software. These software require to be installed on hardware and do not run in the cloud. They also provide a variety of tools and have more complex mathematical modeling in order to get a deeper analysis of the risks and possibly forecast them. These software are fundamentally different from Inclus because they are designed to be more comprehensive and focus more on the modeling of risk.

An old but quite similar to Inclus is the Resolver ballot risk management software. The experts can anonymously give Resolver Ballot information about the risks via voting and the software will then use pairwise comparison to rank the votes. It will also create a heat map visualization of the results

with the standard deviation. Although limited in functionality the Resolver Ballot seems like an early version of Inclus because it has the same fundamental principles.

Thinking portfolio is a strategic planning tool but it also has comprehensive risk management tools. It is hosted in a cloud and can be accessed via browser similarly to Inclus. Thinking Portfolio can be used to elicit information through questionnaires and it also creates graphical representations of the results. Thinking portfolio also has got more functionalities and they are not as focused on just risk management as Inclus is.

4 Methods

4.1 Three-point estimate

Python was used for all three-point estimate related implementations. A function was created for generating mock data. Fitting the triangular and PERT distributions to mock data was very straight forward as both distributions have closed form formulas for using three-point estimates for fitting the distribution. For fitting the GTSP distribution, Algorithm 1 was used.

There were two points to keep in mind while creating the demo implementation. From Inclus' point of view, the computational intensity should stay at a minimum as the software needs to be light enough to work even on relatively old phones. The second point is that as the software is running on JavaScript, we should avoid using complex libraries as shortcuts as they might not be supported in Inclus' framework.

We estimate that the computational intensity should not produce problems as, besides plotting, light data manipulation and mock data generation, the only computations required are to repeatedly solve the equations described in Algorithm 1. In our testing, the algorithm converged within 5 iterations for $\epsilon = 0.0001$. For libraries we only needed to use two more specialized functions: from `scipy.special` we used `sc.beta` to import the Beta function, and from `scipy.optimize` we used `fsolve` to solve the root of the cumulative distribution function of the GTSP distribution. We estimate that these functions are common enough in other programming languages libraries that we decided not to create our own implementations for these. The code is available for review in the appendix, for the curious reader.

4.2 Risk dependencies

4.2.1 User interface design

Currently Inclus is using a questionnaire to elicit information on the cross-impacts of risks. In this questionnaire each risk is a question and the user needs to choose which risks are related and then choose the strength of the relation from a drop-down box. This questionnaire has some strong and some weak features. The downsides are namely the monotonic elicitation process: even though functional, the questionnaire format itself can do little to activate the user through-out elicitation process. The traditional alternative is to fill a N time N matrix, where N is the number of risks, which is not very appealing for obvious reasons. Inclus's form lets user to skip the irrelevant cells but is still text-based. Also the questionnaire does not elicit the sign of the correlation. This is not very problematic when used as a qualitative tool, but including this feature would elicit more useful information for later usage. A good feature in the tool is the reporting of risks as a graph which visually depicts the interrelations of risks. We liked this way of communication and this served as a starting point for our suggestion.

Before starting to design the interface we had an interview with Inclus's client to get better insights to the needs of the user. On the cross-impacts the client did not see much gains for their line of business. Currently the questionnaire used by Inclus has too many things and the end-user client did not find that analysing these cross-impacts would currently bring value to their risk assessment process. They have, however, indicated that for a more quantitative Monte Carlo simulation analysis for these cross-impacts could be useful. At the moment Inclus's tool is foremost a qualitative tool and doesn't offer such analysis. There is some interest by Inclus to potentially develop and extend their tool to this direction. Before the interview we had had similar ideas about the potential development and designed our approach with possible later extension to a more quantitative tool in mind.

We came up with an idea of completely visual tool reminiscent of a mind-map that would offer the user an intuitive way of recognizing cross-impacts when compared to a questionnaire or a N times N

cross-impact matrix. The web-interface would present the user with field where the recognized risks would be shown as nodes on the interface. The user would be invited to draw the arcs connecting the nodes using mouse and keyboard as input. The user would eventually connect the risks into a full graph, which would hold all necessary information about the cross-impacts of risks. We hope that this kind of visual tool could offer more pleasant user experience and work more efficiently at eliciting the interrelations of risks. By eliciting a graph we are implicitly constructing a Bayesian network, which allows for some interesting extension possibilities that are discussed in detail in the next section.

We would like to note, that correlation between risks can be often encountered in literature and is commonly used as a term. However, in this report we use the term cross-impact for this purpose as this emphasizes the causality between risks over the association. This term is also used for example in [10].

When drawing arcs between the nodes on the interface, the user would be prompted for the direction of the causality: uni-directional meaning that the occurrence of risk A affects the likelihood of risk B occurring - but not the other way around - whereas bi-directional causality arcs mean that both risks affect each other. An example of a bi-directional dependency would be a personal injury on a construction site and project delay: an injury could cause delay and delay itself can cause rush and expose workers to hazards that would normally be mitigated but with time-stress might be neglected. Thus the arcs themselves would record the direction of causality.

One important thing about the user interface design is what to show and what to hide. As mentioned earlier, the end-user client's representatives found that the current correlation analysis had too much in it. This further highlights the common design principle about showing only the minimal necessary to the user.

What each user find necessary may differ from user to user. Also as we are only at the stage of designing the interface, we do not yet have definitive information about what is necessary and many options may need to be tested. To address these problems we suggest to have a side panel, with multiple show/hide and highlight options available to the user.

The side panel should contain a slider with which user could control the threshold level for the shown risk nodes. This should be similar to what Inclus's current tool has for result presentation which lets user to see only the cross-impacts that are above the threshold. This would allow the user to explicitly control the level of accuracy of the analysis. Similar slider could be considered also for the risk probabilities and we note this as an additional option.

The side panel should provide means to control for the coloring and other attributes of the interface. Meaningful attributes for controlling are the risk groupings, risk severity/impact, risk probability and the sign of the cross-impact (consistent/inconsistent). Visual means of depicting these include the coloring, size and shape of the objects.

The sign of the cross-impact could be displayed by coloring the user-drawn arcs with a two-sided color scale: for example green could indicate that the risks are consistent (A happening increases the probability of B happening) and red the opposite. Also the color opacity could be used to further highlight the strength of the cross-impact, but this is not necessary as the arc size already shows this, and this would be redundant.

The risk severity/impact and probability could be visualized by coloring the background of the risk nodes. This could be done with a single color where the color opacity would be proportional to the impact or probability. At least for risk impact red would be natural choice. The same attributes could be illustrated by the size of the nodes. We suggest that the user should be presented with a drop-down list for both color and size options.

How to show the risk groupings - or not to show - requires more contemplation. Initially we thought about coloring the background of the interface in which the risk nodes are plotted into separate regions by the risk category. We noticed from our own project risk assessment for this report that risk groupings are always a matter of taste and certain risks can have several suitable categories, and deciding over this can be arbitrary. Thus coloring the background can direct user's lines of thought too much, though we still note this as one option even if it is not our first recommendation. This is illustrated in the Figure 9. Second option is to color the risk nodes with colors corresponding to their categories. This is illustrated by in Figure 10. Coloring the nodes has the drawback that the color attribute might be more suitable to visualize other attributes, like risk probability, and thus not available. Third option is to plot the shape of the nodes according to the categorization. Possible forms are ellipses, squares, triangles, etc. We cannot yet suggest the most meaningful default setting for coloring, so this should

be empirically tested when the interface is implemented.

Another way to describe the risk groupings which is independent of the above mentioned ones is to plot the risk nodes initially in clusters so that the physical distance on the screen is small for the risks within same category and the distance between these clusters is larger than within-cluster distances. The proximity gives natural implication that the risks belong to the same category. We also presume that the risks in same groups might have the most cross-impacts, which means that plotting the nodes close to each other eases the user's task of recognizing cross-impacts. We also do not believe, that this kind of loose clustering biases the user's lines of thoughts too much. We recommend plotting categories as loose clusters as a default view. If more emphasis is wanted this can be can be combined with the three possibilities mentioned before. We also note that the risk nodes on the interface should be movable with click-and-hold method so that the user can reposition the nodes as most preferred.

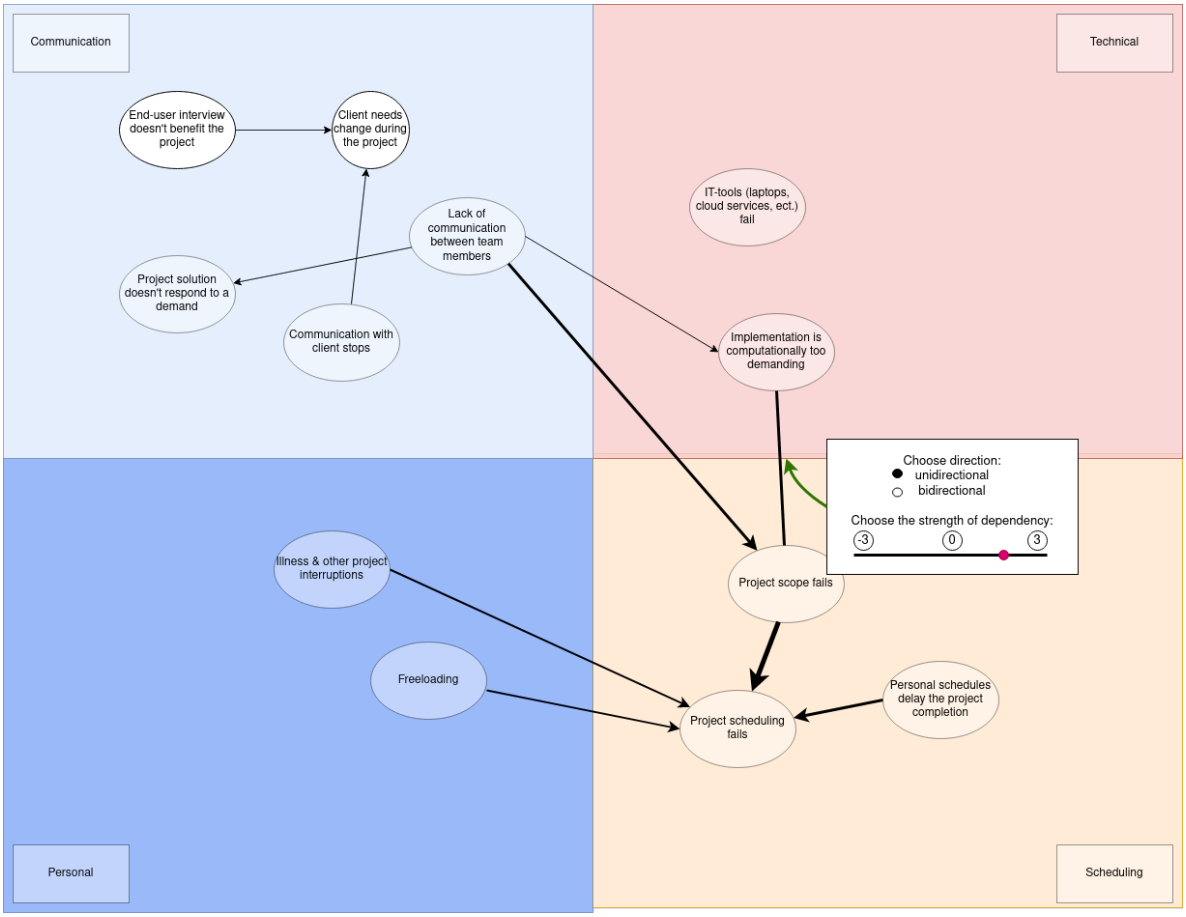

Figure 9: An illustration of the suggested tool. The background is colored to illustrate the risk groupings. The side panel is omitted.

Drawing the arcs should be made as easy enough for the user. We recommend that the user is allowed to select risk nodes with left-click and that the software automatically starts drawing an arc from the node when selected without explicit command thus minimizing the amount of necessary clicks. User then would select the end-point with another click and would have drawn an arc with just two clicks. After this a pop-up-window would appear automatically next to the drawn arc, prompting the user to define the details.

The pop-up-window would ask about the direction of causality. The default option should be that the first selected risk node impacts the second node. This could be changed to reverse direction or bidirectional arc by clicking the corresponding option as is seen in Figures 9 and 10. Symmetrical direction would mean that both risks can affect the occurrence probabilities of each other. In bi-directional case the cross-impacts are assumed to be identical even if in reality this might not be the

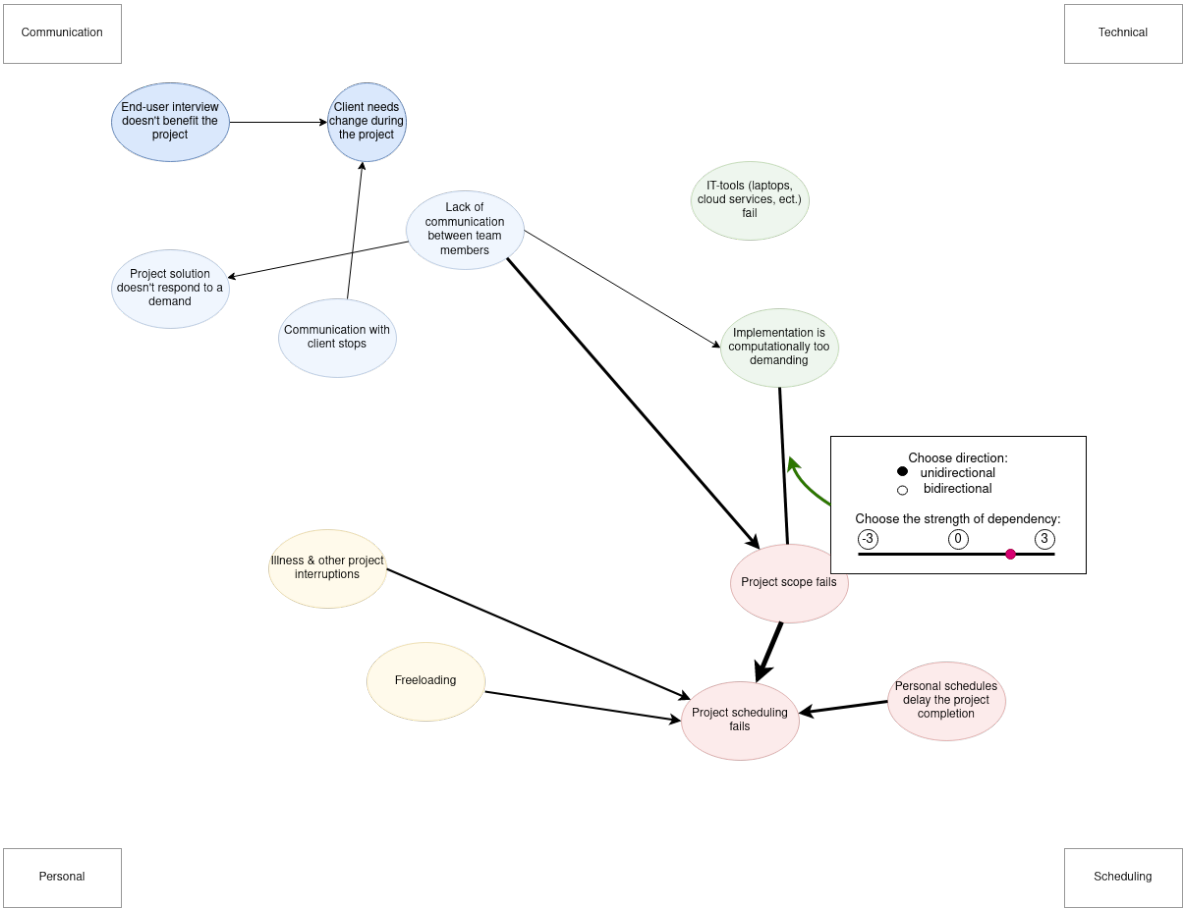

Figure 10: An illustration of the suggested tool. The risk groupings are depicted with the node colors. The side panel is omitted.

case. This assumption should not simplify the analysis in a way that causes negative impacts on accuracy - after all the first goal of the tool is to serve as a starting point for further analysis and facilitate risk management conversation.

After the direction of causality the sign of the cross-impact is determined. There are more than one possible way to elicit this, but we recommend the use of Likert-scale for this as this approach is already used by Inclus and is thus consistent with other parts of Inclus software. Specifically a 7-point Likert-scale ranging from -3 to +3 has been suggested in the academic literature [23]. In this scale negative values mean that the risks are highly inconsistent or exclusive and one happening decreases the probability of the other risk occurring. -3 would correspond to completely exclusive/inconsistent. Zero would mean that the events are independent and do not affect each other's occurrence probability. Positive values mean that the first risk occurring increases the probability of second risk occurring and the risks are thus consistent.

This symmetric Likert-scale is easy and natural for the user, but it has the downside that it is not directly convertable to real numbers for numerical analysis, as probabilities must be non-negative values and the probability ratio is not bounded. Zero in Likert-scale would be converted to 1 in probability ratios as 1 stands for independence in ratios. -3 would be converted to 0, meaning strict exclusiveness. Choosing what +3 stands for is harder to decide on. Some possible candidates would be 100, 1000 and 10^{10} where the latter one would be numerical representation of perfect causality. Also suitable numerical conversions for numbers between zero and the minimum and between zero and the maximum must be decided.

Another possible scale worth considering is 0-5 and a separate selection of the sign of the impact. First the user would be prompted to select whether the the connection is increasing or decreasing the

probability. Suitable default setting would be increasing, as we expect based on our own experiences this to be more frequent connection. Next the strength would be elicited on the scale from 0 to 5. This would allow more accuracy and better use of the whole scale as the bi-directional scale is cropped to its meaningful part. We note both options: a single two-sided scale and two-phase selection, where the scale itself would be one-sided, and leave for the client to choose which is more preferable.

There is also the possibility of eliciting a probability distribution for the cross-impact instead of a single point estimate. This could be done by the three-point estimation method presented earlier in this report. Then the single point slider would simply be replaced by the three-point estimate interface on the pop-up window.

After the pop-up window is fulfilled with the needed information, an arrow would be drawn connecting the nodes. The head of the arrow would show the direction of causality and the thickness of the arrow the strength of the relation. The color of the arrow would show the sign of the cross-impact, e.g. green for increasing the probability and red for decreasing. Other suitable color pairs can also be used. We note that there may be needs to later adjust the stated cross-impacts, so the user should be allowed to change the parameters by later clicking the arrow, when the pop-up window would reappear allowing for modifications.

We suggest that the user be presented the option to see additional information of each risk elicited earlier in the Inclus's workflow by right-clicking the wanted risk node. A pop-up window showing the risk's grouping category, the earlier estimated probability and severity as well as possible additional notes could be useful for the user. Alternatively this information could be automatically shown on a separate side panel on the right-hand side of the screen. However, we note that this decreases the space available for the main panel, and although possible, we do not include this in our recommendation.

The user may faced with the problem of the interface being over-crowded with large number of risks. To counter this we suggest that the view has the possibility of zooming in/out and panning by clicking and holding an empty spot of the canvas. The side panel could also have the option of hiding certain risks based on user chosen criteria like risk grouping. We leave the decision whether this hiding feature is necessary to the client.

Potential pros of our approach are the more intuitive experience offered by a visual tool. We expect our suggestion to offer more pleasant user experience than questionnaire could provide. One of the strong sides is also that the user sees the end-result forming and may stop when they feel the elicitation process is finished. As the user interface is the final result at the same time, we expect there to be less inconsistencies and surprising results that could occur with questionnaires when the user looks at small details at once missing the big picture.

We also note some possible problems with our approach that should be known. First of all, the visual tool is likely to be much more difficult to implement than a questionnaire, and may require considerable amount of work from the developing team in comparison. Also with many risks the graph layout may look rather crowded and be hard to use. For this reason we have suggested the slider to show only the risks that are above certain threshold as well as the option of focusing on certain group although these do not completely solve the issue if the number of risks is relatively large.

We suggest that in the aggregation phase the joint results from multiple experts could be reported with a similar graph structure, but instead of showing the cross-impact strengths and other information, only the arcs would be shown. The opacity of the arcs would present the portion of the experts that recognized the interdependency, where almost translucent arcs would be rarely recognized interdependencies and thick arcs often recognized ones. Another way to present would be the thickness of the arcs instead of opacity, but we do not recommend this alternative as there is possibility to mistake it for the strength of relation as this way of plotting is done for individual experts. Again a threshold slider displaying the portion of the respondents would be useful. This graph would show the scatter of the elicited expert opinions.

One way to aggregate the different graphs to one is to use a significance threshold for arcs: for example 25 % of the experts need to recognize the connection before it is included in the results. The arcs included could then be aggregated by taking the average of the estimated cross-impacts.

In this approach significant inconsistencies may appear and they need to be recognized. Example of inconsistency would be arcs with opposing directions given by two different experts. When there is clearly a dominating direction and only small number of opposite arcs the problem is resolved simply by treating the arcs as separate arcs (which they are) and the threshold condition eliminates the dominated arc. When both directions have significant support the problem is more demanding. Our

suggestion is to flag the arc as needing further inspection and the matter could be discussed later jointly with the experts and resolved.

Other inconsistencies will likely appear, making straight-forward aggregation impossible. Providing a systematic way of aggregation requires further information on the arising inconsistencies. Thus we cannot yet provide a systematic approach for this and leave deciding on the aggregation method for the client.

4.2.2 Further development

The user interface was described in the previous section and when implemented this should give a sound alternative to replace Inclus's current tool, which was one of our main objectives. Besides suggesting replacement for the current tools, we also seek to provide advice on possibilities on extending Inclus's tool to tasks to which the tool is as of yet unsuitable. On this section we describe these ideas and methods which allow the extension of the introduced mind-map-like cross-impact elicitation.

We suggest analysing the created graph in qualitative manner. When the user is finished building the graph and has submitted the work, the software could analyze for example which nodes are isolated, i.e. they are not connected to other nodes by any arcs. Another interesting result to report is which risk nodes are the root causes, i.e. the starting points of arcs, but no arc ends there. Similarly end nodes could be recognized. Also risk nodes with exceptionally many arcs originating from them could be reported, as those risks could have several consequences. Some threshold for these statistics could be considered: risk nodes with small base probabilities, small cross-impact effects and small severity may not be worth the user's attention in the presence of far more severe risks. We believe this kind of descriptive statistics could be very useful for the end-user. Also implementing a piece of software program to recognize these graph characteristics should be rather straight-forward task.

Another interesting result for reporting could be drawing tentative fault-trees for high-impact risks, for example for risks with severity over 4 in 0-5 Likert-scale. These are the big "end-risks" which the user is likely most interested in preventing and thus highlighting them by presenting them separately might be useful. This is foremost a question on reporting the most essential findings of the analysis. Also we note that these presentations are not comprehensive fault-trees found in risk analysis literature and other tools, and the user should be informed not to treat them as such but more as starting points for deeper analysis.

One of the major limitations of Inclus's tool is its lack of quantitative analysis. Now, with the suggested mind-map-like approach this is possible, as the end product the user is making is kind of Bayesian network with qualitative descriptions. When the qualitative Likert-scaled statements from both the probability and cross-impact elicitation parts are converted to appropriate numerical values and suitable distributions are introduced we get a proper Bayesian network. Then numerical analysis can be carried out if needed.

The network requires the occurrence probability to have been divided into *base probability* and *cross-impact posterior*. Base probability would mean the probability of occurrence given that there is no knowledge on the occurrences of other risks that can affect its probability. Cross-impact posterior would be the effect caused by other risks occurring. The base-probability would equal to the three-point estimates elicited and the cross-impact part would be the weights of the mind-map-like graph. We note that in reality this separation might not be so easy: the user would need to be specifically asked about the probability *given that* no other risks have occurred. In complex case this might not be possible to even observe. Thus the user might easily think of the joint probability instead of the base-probability, or in other words: there might be leakage from joint probability to base probability. This needs to be considered if numerical probabilities are to be presented to the user.

Suitable distribution for the most risks is the Bernoulli distribution

$$f(k,p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$

which gives binary outcome as a result. If the risk can occur several times binomial distribution could be used. In this report we are going to demonstrate only the Bernoulli distribution. The prior for the p would be the formerly elicited three-point estimate converted to numerical probability distribution. The cross-impact multipliers would be the arcs elicited in the graph. From the cross-impact multipliers

the posterior probability of another risk can be calculated with the formula

$$P(B|A = 1) = P(B) \cdot C_{AB}$$

where $A = 1$ means that risk A has occurred.

When the cross-impact multiplier is assigned a distribution as well instead of a single point estimate, we get a two-dimensional joint distribution of risk B and cross-impact multiplier. Then sampling can be done from this joint-distribution. If simpler calculations are wanted only the mode can be used as well, when the distribution becomes one-dimensional again.

Now we have a fully functional Bayesian network ready for quantitative analysis. One interesting analysis possibility we suggest is a Monte Carlo simulation for the network. In this analysis N instances of the constructed network would be initialized with the base probability rates and cross-impact multipliers. Then according to the base rates occurrences would be generated, or risks would be "triggered", over the lifespan of the case in question. When a risk occurs, the connected risk probabilities would be updated and the simulation would continue until the lifespan end would be achieved, thus generating a possible case scenario. Finally the results would be aggregated into statistics how each risk occurs. When combined with the elicited risk impact data we could calculate the expected risk harm alongside with its distribution. These statistics are often very valuable to the client and the focus of many other risk analysis software.

The posterior probabilities considering the cross-impact effects can also be calculated analytically when the base rate distributions are analytically defined. This would also be computationally cheaper. This is of course impossible if the distributions given were to be not analytically defined, which would require Monte Carlo sampling. Another positive side on the Suggested Monte Carlo method is that *chains* of occurrences can be found. This way we gain information on the possible development paths of events - an information that is not included in the analytical results. When these chain reactions are recognized we can automatically recognize potential intervention points. For example, if a path of three risk nodes and two connecting arcs shows up, where the base of the chain is a common low-impact risk, but the end of the chain a high-impact event, we could direct our attention to the first risk node, either so that we prevent the risk from occurring at all or try to mitigate the effect that the risk occurrence has on the other risk nodes' probability. Intuitively we believe that developing an algorithm for this path and intervention point recognition should not be overly complicated task, although we leave presenting such an algorithm outside the scope of this report.

We recommend that when this kind of Monte Carlo simulation is performed, that the updated posterior probabilities would be shown on the graph. On a separate graph the risk paths should be visualized along with the potential intervention points. This would provide a new basis for discussion and further risk management planning that Inclus does not as of yet provide and which might be of significant interest to clients. Once more, however, we would like to draw attention the aforementioned information leakage from the joint posterior distribution to the base probability. If great care is not put in how to communicate and elicit the base probabilities and cross-impacts the numerical values provided by the suggested Monte Carlo simulation can be inaccurate and thus we recommend this analysis to be used only as guiding results. In other words, the results should be reported qualitatively, only to facilitate further analysis. No critical decisions should be based on the numerical results without sufficient further research to the suggested method.

One possibility that may potentially hold interest is exporting the generated Bayesian network in code format for example in Stan, Python or other software that allows analysis of Bayesian networks. This is done in the commercial software BayesiaLab which provides a visual tool for working with Bayesian networks. This kind of easy exportability could be appealing to organizations that have the capabilities to work with quantitative risk analysis, but use Inclus's tool mainly for recognizing and discussing risks.

5 Results

5.1 Three-point estimate

Having already shown the very basic behaviour of all three suggested probability distributions, on this section we focus more on creating points of comparison to analyze how the distributions differ from each other in different situations. We will also take a look at how the distributions look when the opinions of multiple experts are aggregated and create comparisons for these as well. The range of possible values was limited in the following Figures between $[0, 1]$ for simplicity's sake. However, any arbitrary range can be used with any distribution by the use of normalization.

Figure 11 shows the probability distribution functions for a symmetric three-point estimate of $[0.3, 0.5, 0.7]$. For PERT distribution a weight of $\gamma = 4$ was used. For GTSP distribution quantiles $p = 0.025$ and $r = 0.975$ were used. We can see that GTSP has the highest mode while also providing the longest tail with the most probability mass. This means that, for symmetric quantiles with the selected parameter values, GTSP gives the most weight to values near the mode and the most probability mass to values near the selected extreme values. Comparing the triangle and PERT distributions, we can see that the triangle distribution has more probability mass on the tails and less on near the mode. In other words, PERT distribution gives more probability mass to values near the mode and less to extreme values, compared to the triangular distribution.

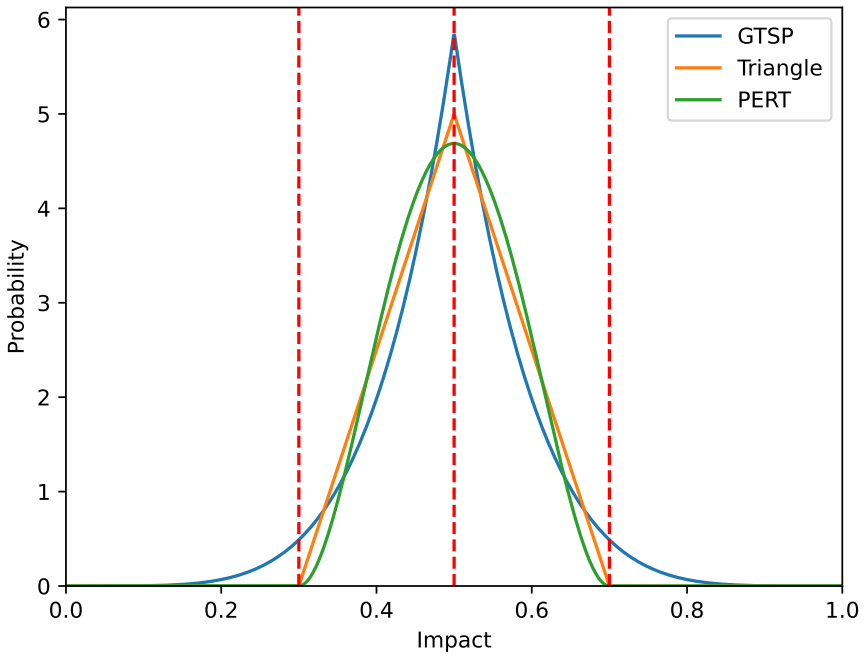

Figure 11: The probability density functions of the GTSP, triangle and PERT distributions for three-point estimate $[0.3, 0.5, 0.7]$.

The behaviour of our candidate distributions for asymmetric three-point estimates is also of interest, as we suspect that these will be encountered more in real life. To this end, we created two graphs for comparison, in Figure 12 slightly skewed three-point estimate of $[0.3, 0.5, 0.55]$ was used and in Figure 13 very skewed three-point estimate of $[0.1, 0.5, 0.55]$ was used.

From Figure 12, we can see that the same analysis mostly applies here as was used with the symmetric three-point estimate. The GTSP distribution has the highest mode with the most probability mass at the tails, PERT distribution has the second highest mode and less probability mass at the tails compared to the triangle distribution. However, the slight skewness of the distributions makes the differences in the left-side tail more pronounced. The left-side tail of the PERT distribution has clearly the lowest probability mass, implying that with skewed distributions PERT distribution has

the tendency to give very low probability mass to long tails. The triangle and GTSP distributions, on the other hand, both have significant probability mass near the extreme points. GTSP distribution tends to give less probability mass between the mode and extreme points, compared to the triangular distribution. This is due to the higher mode and longer tail of the GTSP distribution.

In Figure 13, the extreme skewness of the distributions make the previously mentioned differences clear. Because the PERT distribution gives small probability values for long tails, the more skewed the distribution the more probability mass the mode tends to get compared to the other distributions. Also, with the long left-side tail, the probability values around the general area of the minimum value are on a different magnitude compared to the GTSP and triangle distributions. This, however, could be corrected by choosing a lower value for the weight parameter $\gamma = 4$, that would give the tails more probability mass. Changing the value to $\gamma = 2$, we can see from Figure 14 that the PERT distribution once again bears a resemblance with the other distributions. The GTSP and triangle distributions resemble each other with the difference, that the GTSP distribution has longer tails on both sides, leading to less probability mass near the mode.

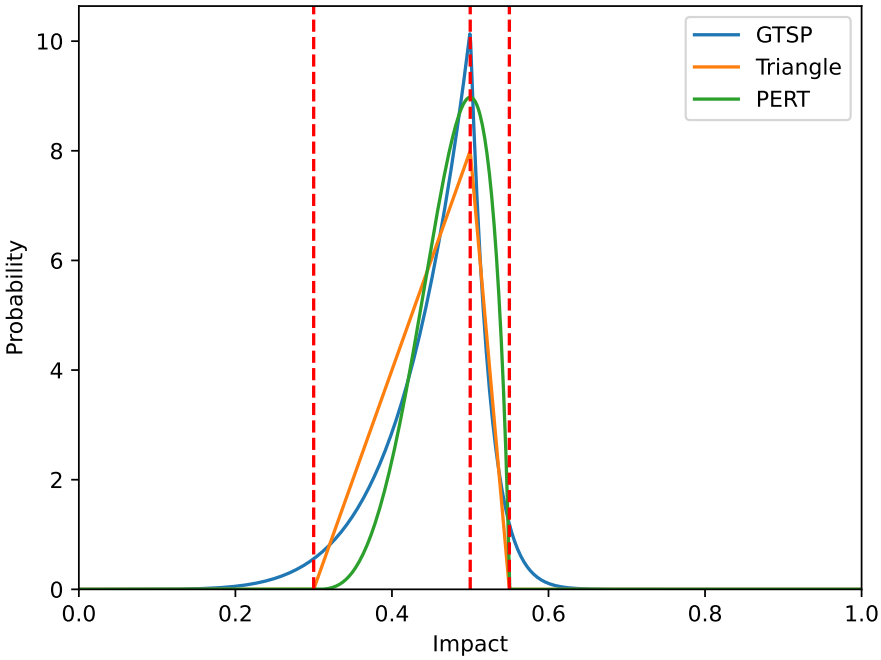

Figure 12: The probability density functions of the GTSP, triangle and PERT distributions for three-point estimate $[0.3, 0.5, 0.55]$.

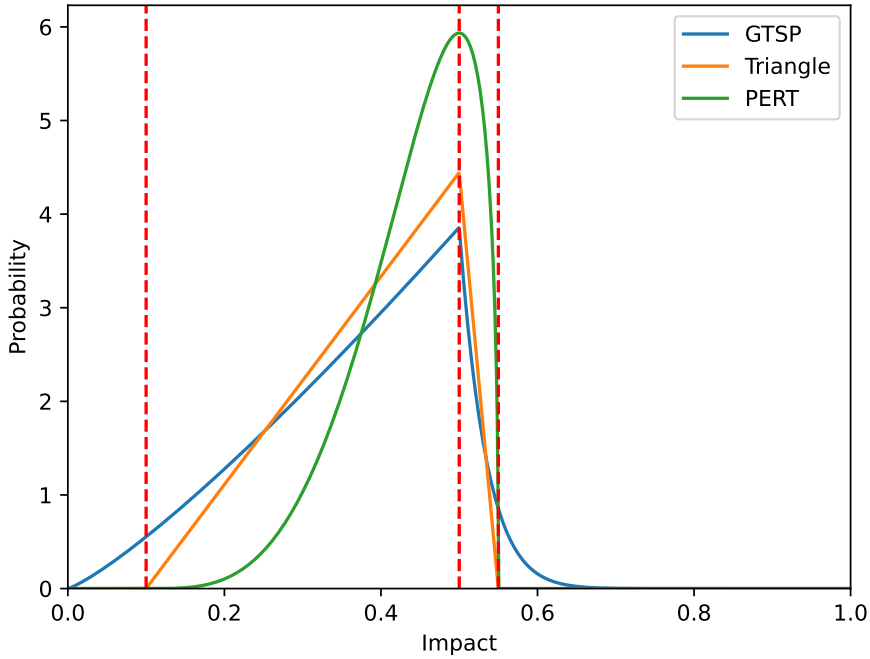

Figure 13: The probability density functions of the GTSP, triangle and PERT distributions for three-point estimate $[0.1, 0.5, 0.55]$, with PERT parameter $\gamma = 4$.

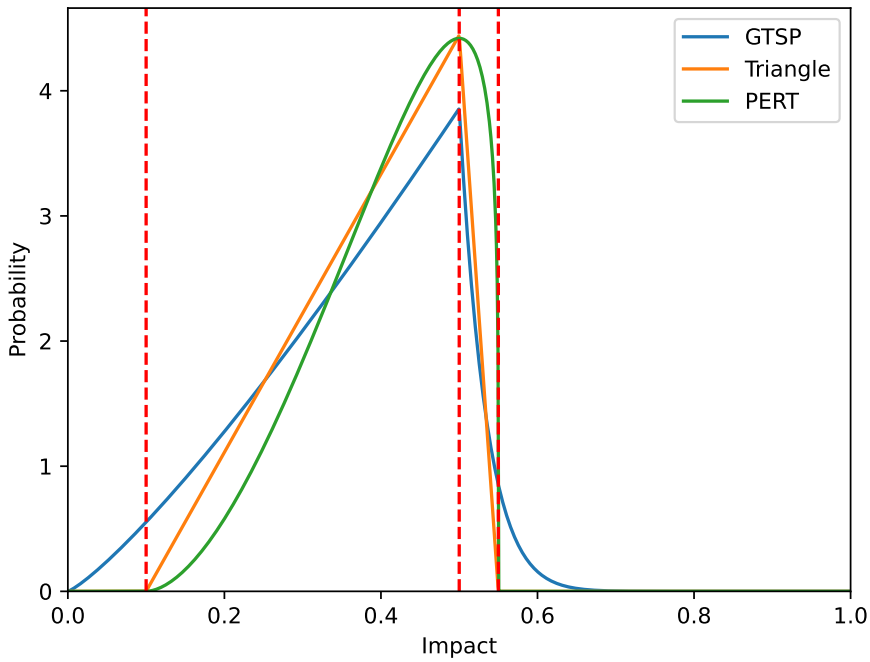

Figure 14: The probability density functions of the GTSP, triangle and PERT distributions for three-point estimate $[0.1, 0.5, 0.55]$, with PERT parameter $\gamma = 2$.

5.2 Aggregation results

Next, we look at these distributions, when the opinions of several experts are aggregated. For simplicity, we used a weighing scheme where each opinion had the same weight. Figure 15 explores how the aggregated distributions behave with three slightly skewed but similar opinions. Figure 16 explores how the aggregated distributions behave with differing and slightly skewed expert opinions.

Several of the previously encountered phenomena are present in Figure 15. The GTSP distribution is flattest out of the bunch due to 5% of the probability mass being outside of the minimum and maximum values. At the other end of the spectrum, the PERT distribution has a very sharp peak compared to the other distributions. Here, again, we can see that the PERT distribution gives the tails little probability mass relative to the other distributions. The triangular distribution is between the GTSP and PERT distribution with regards to the height of the mode and the probability mass found in tails.

When experts have slightly more differing opinions, as presented in Figure 16, the shape of the distribution near the mode has a big effect on the aggregated distribution. The GTSP distribution has two peaks that are around the same height, with the probability decreasing between the peaks. The other distributions, on the other hand, have a clear maximum value even though their left- and right-sides are not monotone looking from the mode.

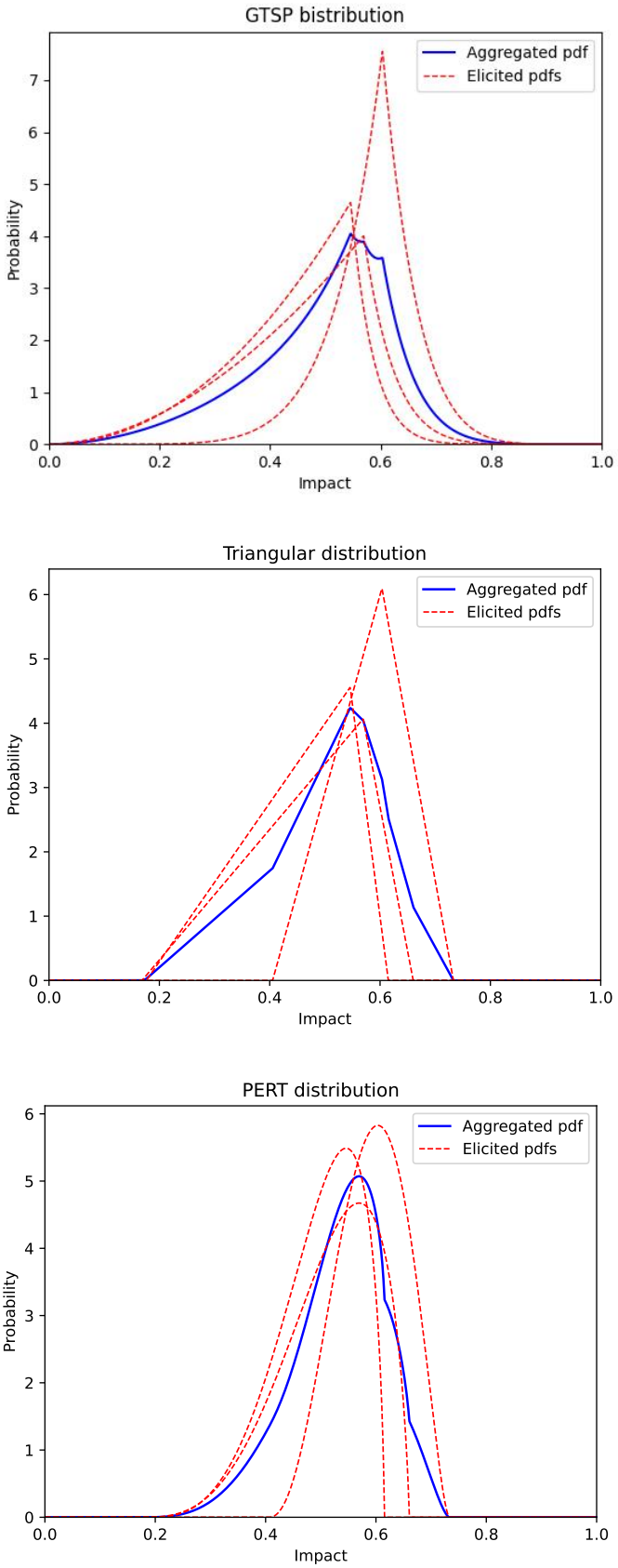

Figure 15: Aggregated distributions with similar expert opinions.

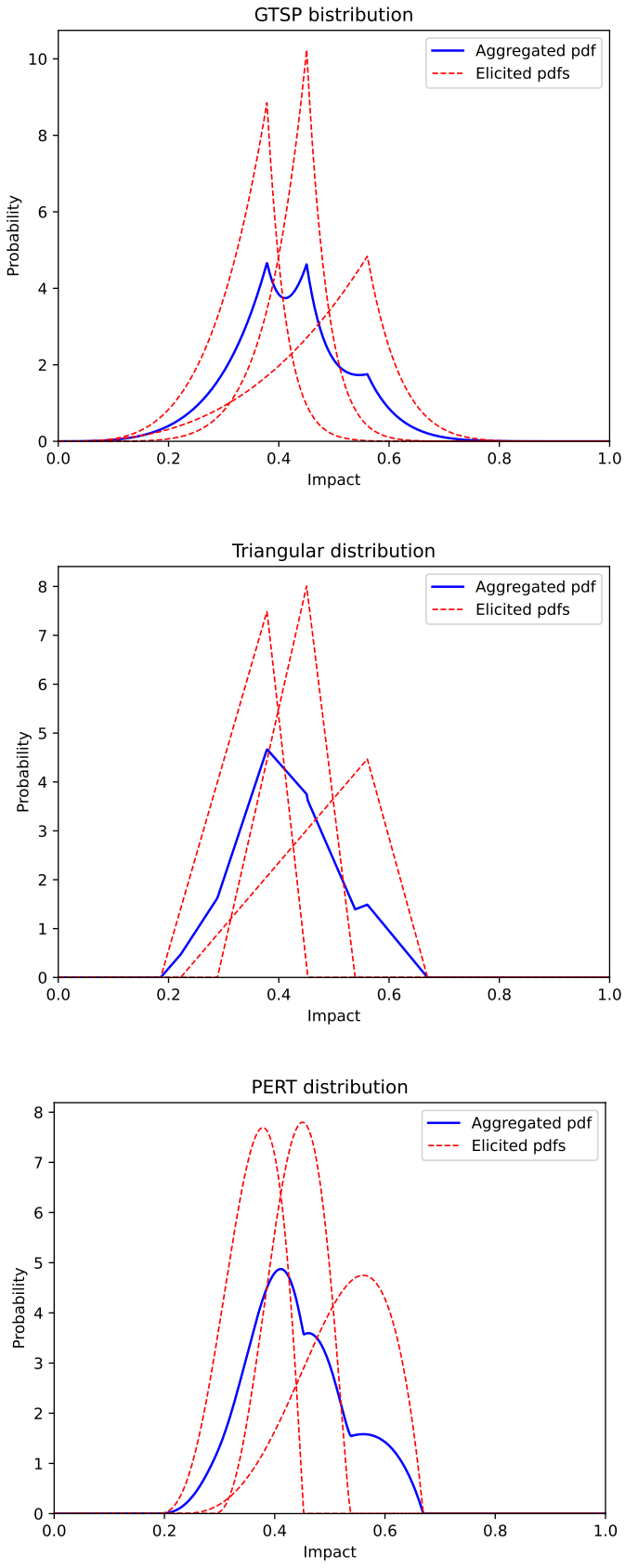

Figure 16: Aggregated distributions with differing expert opinions.

The characteristics of these three different distributions can be utilized in choosing the method in IDEA protocol. PERT distribution and triangular distribution limit the probability mass strictly between the lower and upper bound. This is suitable for three-step elicitation in IDEA protocol. In GTSP distribution, the probability mass extends past the given bounds, because they are defined as quantiles. We can apply IDEA four-step elicitation directly to this by taking the standardized credible intervals and using the as parameters in creating GTSP distribution.

Since the final distributions to be displayed were calculated directly as a weighted sum of individual distriburions, some perturbances and unnatural discontinuities in the slope of the density function can be perceived. They occur at the maximum and minimum given values of individual distributions if PERT or triangular distributions are used. If GTSP is chosen, there are spikes in the given mode values. However, the exact values of individual estimates of mode, lowest and highest values are irrelevant for the true aggregated distribution. It is more important to capture the general behavior and trend in estimates of experts. We can assume that the inputs of experts follow some distribution such that the aggregated distribution converges to the actual distribution when the number of experts grow large enough. It is also reasonable to assume that the actual distribution should grow monotonically between the minimum value and the mode, and decrease monotonically between the mode and the maximum value, perhaps resembling a behaviour of normal distribution. The sudden changes in slope in the aggregated distributions should be smoothed to get a closer match to the true underlying distribution. Smoothing is further investigated in the next chapter.

5.3 Smoothing

One simple smoothing method is applying nearest neighbor smoother. This is performed by first choosing a discretization interval $1/n$ and discretizing the PDF. Then we choose a number k to denote the range of calculating the average. It describes how many neighboring values in both sides are taken into account. The discretized PDF can be padded with extra zeroes, when the normalized impact value is smaller than zero or larger than one. This enables moving average values at the ends of the distribution. If a large k is chosen, the PDF is a lot smoother shape compared to a smaller value of k. If we have a plenty of expert judgement data available with broad upper and lower limits, smoothing is not nearly necessary. However, if the number of expert opinions is small, the credible intervals are tight and the opinions differ from each other radically, large value of k is useful to smooth out the spikes of individual opinions. Figure 17 illustrates the moving average smoothing to aggregated PERT distributions with different modes, lowest and highest values. The value of k/n is set at 0.07 and $\gamma = 4$. Same expert estimates are used in Figure 18, but the chosen distribution is triangular.

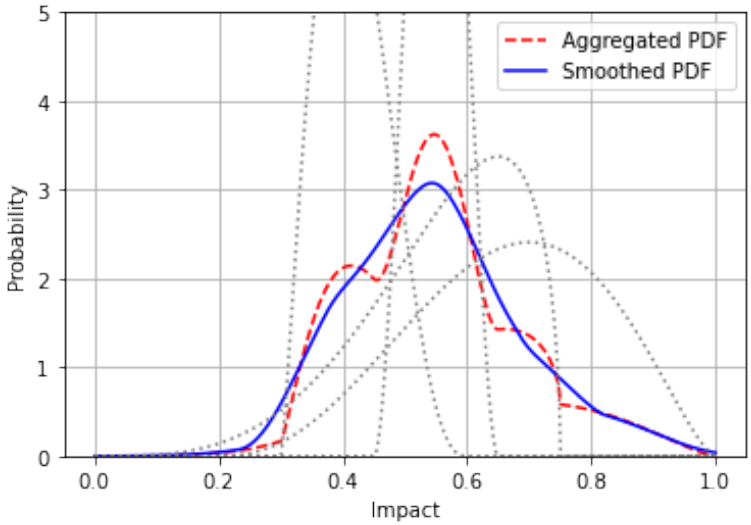

Figure 17: Smoothed aggregated PDF using PERT distributions and nearest neighbor smoother

Nearest neighbor smoother is only one method for performing smoothing and there are other more advanced methods such as convolution methods and Kernel smoothers in general. Kernel smoothers

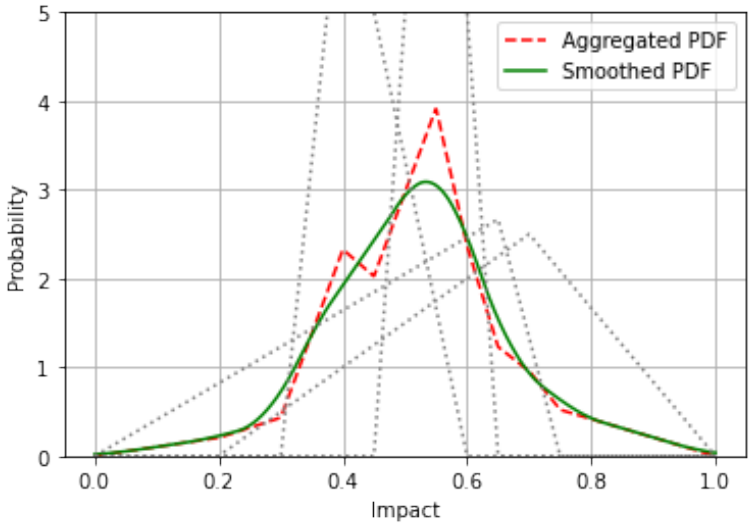

Figure 18: Smoothed aggregated PDF using triangular distributions and nearest neighbor smoother

estimate the function as a weighted average of neighboring data, and nearest neighbor smoothers is a subcategory of Kernel smoothers where data is weighted equally.

The smoothed probability distribution function can be denormalized to the original scale and displayed as in Figure 19. It is an example where opinions of 4 experts and chose triangular distribution. Nearest neighbor smoother was applied to capture the underlying PDF. Quantiles and expected cost are visualized in the image. Any quantiles can be obtained from the cumulative distribution function and expected value can be calculated as a numerical integral or a weighted average. The PDF can be for example use to calculate the probability that the costs exceeds a given threshold and expected values conditioned to threshold exceedings.

6 Discussion

With regards to the practical implementation of a three-point estimate, there are several key questions that should be asked when deciding what distribution would work the best.

From a questionnaire standpoint, we want to allow the experts to choose the minimum, most-likely and maximum values as freely as possible. This can give rise to very skewed distributions, where the PERT distribution could suffer as the weight parameter γ is not something that the end-user would most likely choose. Although, the skewness would depend on the specific support used in the questionnaire as well. Choosing the parameter could also be left as an extra option in the settings for the mathematically trained individuals.

The triangular distribution gives much more weight to values between the mode and extreme values compared to the other distributions. We suspect that the shape of the distribution of an impact would bear resemblance with normal distribution in the sense that most of the probability mass would be near the mode and tails would have relatively little probability mass. From this point of view, the GTSP or PERT distribution should be favoured. However, this is just an assumption as the true distribution is unknown. The GTSP distribution, on the other hand, has a tendency of generating multiple maximum values in the aggregated distribution for differing opinions due to its sharp shape near the mode. This could lead to confusion when end-users try to interpret the results.

The rapid changes caused by individual opinions can be observed easily in the aggregated distribution. Smoothing can be used to make these perturbations vanish and to capture the underlying behaviour. Smoothing also makes the results more intuitive and interpretable.

Using different distributions for different risk-assessment projects of the same client should be handled with care, as inconsistent results between projects can easily lead to confusion. The results should be intuitive and consistent.

One big limitation of our study is that we were not working with real data nor with real customers,

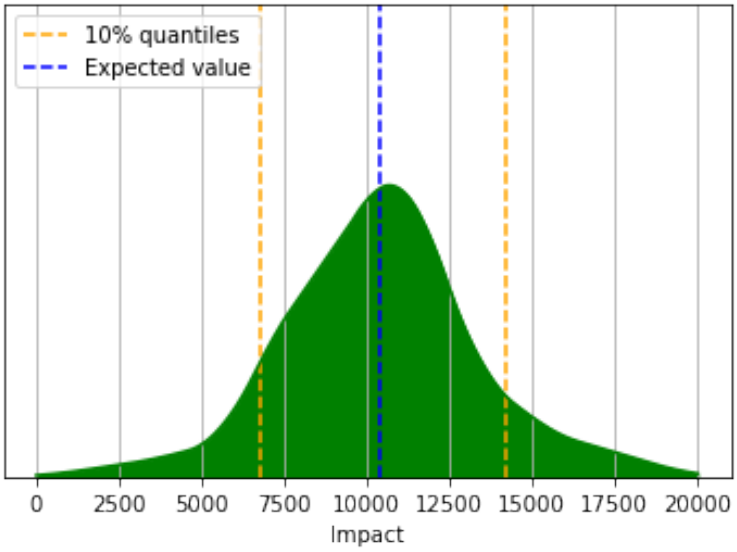

Figure 19: Smoothed aggregated PDF using triangular distributions and nearest neighbor smoother. Impact scale is denormalized. 10 percent quantiles and expected value of the financial cost of the risk is shown in vertical dashed lines.

so much of the analysis is based on purely our own intuition. We recommend experimentation with all three distributions in a real risk-assessment project, keeping in mind the possibility of changing the quantiles of the GTSP distribution and the weight parameter of PERT distribution. Clients would surely have invaluable feedback on the results. It would also be very helpful if historic data was available for comparing how the different distributions would fare against each other.

We provide Inclus' with the code that was written for this project so they can experiment with different distributions using different parameter values to try for themselves how the distributions behave. We encourage working with end-users for gathering feedback on what works and what does not.

7 Conclusions

We benchmarked existing software that are similar to Inclus. The benchmarking focused on user experience and a basic analysis of the methods these software use. Similarities to Inclus were found in all of the software but also distinct differences. The major differences were the required computational resources and domain expertise. The software closest to Inclus were found to be Thinking portfolio and Resolver Ballot. The differences between these software and Inclus are related to the methods of elicitation of information and presentation of results.

We examined a framework for eliciting expert judgements, which is the basis for generating data for three-point estimation. We were able to find and provide analysis for three distinct distributions for use in three-point estimates. The results showed that no single distribution was better than all the rest in all aspects. Some were easier and more intuitive to understand, some fared better with skewed data and some required less attention in choosing free parameter values. Smoothing of the aggregated distributions increased intuition and interpretability, and the produced smoothed PDF can be used to calculate different descriptive statistics, quantiles and probabilities for different uses. Even though the accuracy and complexity of our model might not be at the highest level, this project acts as a good starting point for Inclus to delve deeper in to the world of quantifying probability distributions from expert opinions.

For identifying and analysing risk interdependencies we developed an elicitation approach reminiscent of mind-map. For this suggestion the user interface was documented and illustrative sketches were presented. Also some ideas on how to further extend the Inclus tool for more advanced reporting and also to quantitative analysis were presented.

References

[1] International Organization for Standardization. Iso 31000:2018 - risk management — guidelines - iso. 02 2018.

[2] Henning V. Burgman M. Hanea A. McBride M. Wintle B. A practical guide to structured expert elicitation using the idea protocol. *British Ecological Society*, 2017.

[3] Adams-Hosking C. McBride M. Baxter G. Burgman M. de Villiers D. Kavanagh R. Lawler I. Lunney D. Melzer A. Menkhorst O. Molsher R. Moore B. Phalen D. Rhodes J. Todd C. Whisson D. McAlpine C. Use of expert knowledge to elicit population trends for the koala (phascolarctos cinereus). *Diversity and Distributions*, 22(3):249–262, 2016.

[4] R. Cooke A. Colson. Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety*, 163:109–120, 2017.

[5] Charles E Clark. The PERT model for the distribution of an activity time. *Operations Research*, 10(3):405–406, 1962.

[6] D.G. Malcolm, J.H. Roseboom, C.E. Clark, and W. Fazar. Application of a technique for research and development program evaluation. *Operations research*, 7(5):646–669, 1959.

[7] P. Buchsbaum. Modified PERT simulation. *Great Solutions, Rio de Janeiro, Brazil*, 2012.

[8] Samuel Kotz and Johan René Van Dorp. *Beyond beta: other continuous families of distributions with bounded support and applications*. World Scientific, 2004.

[9] José Manuel Herrerías-Velasco, Rafael Herrerías-Pleguezuelo, and Johan Rene Van Dorp. The generalized two-sided power distribution. *Journal of Applied Statistics*, 36(5):573–587, 2009.

[10] Ahti Salo, Edoardo Tosoni, Juho Roponen, and Derek W Bunn. Using cross-impact analysis for probabilistic risk assessment. *Futures & Foresight Science*, page e2103, 2021.

[11] Christoph Werner, Tim Bedford, Roger M Cooke, Anca M Hanea, and Oswaldo Morales-Napoles. Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research*, 258(3):801–819, 2017.

[12] Roger Cooke et al. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand, 1991.

[13] Alireza Daneshkhah and JE Oakley. Eliciting multivariate probability distributions. *Rethinking risk measurement and reporting*, 1:23, 2010.

[14] Stephen C Hora and Erim Kardeş. Calibration, sharpness and the weighting of experts in a linear opinion pool. *Annals of Operations Research*, 229(1):429–450, 2015.

[15] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[16] Steve Krug. *Don't make me think!: a common sense approach to Web usability*. Pearson Education India, 2000.

[17] Zef.fi. Zef.fi website. https://www.zef.fi/?hsLang=en, 2022.

[18] Workiva. Workiva datasheet. https://www.workiva.com/en-nl/solutions/enterprise-risk-management, 2022.

[19] Archer. Archer operational risk management. https://www.archerirm.com/, 2020.

[20] Vose. Vose webpage. https://www.vosesoftware.com/index.php, 2022.

[21] BPS Resolver Ballot. Bps resolver ballot risk assessment. https://www.bpsresolver.com.

[22] Thinking Portfolio. Thinking portfolio whitepaper. https://thinkingportfolio.com/products/thinking-portfolio-risk-management-portfolio/, 2021.

[23] Roland W Scholz and Olaf Tietje. *Embedded case study methods: Integrating quantitative and qualitative knowledge*. Sage, 2002.

A Self-assessment

A.1 Project plan with respect to the project

A.1.1 Scope

The project finished according to the project plan and there were no major deviations from the initial plan. The planned sections were finished as discussed with the client. Some minor re-weighting happened during the project, but this impacted the end result relatively little.

A.1.2 Risks

Our initial project risk assessment proved rather accurate and there were no unexpected risks. During the later stages of the project personal scheduling risks realized, which was not unanticipated. No other risks with negative impacts were observed. We had also discussed counter-measures for the realized scheduling risks and there was no specific negative impacts.

A.1.3 Schedule

The tasks in the original schedule were very generic. As we did not have specific schedules for specific tasks, following the original schedule was very easy. In hindsight, we could have made the original schedule more detailed.

A.1.4 Project execution

Project execution followed the project plan as stated before. Different parts of the project were divided between the members of the team and they were executed in parallel. The team held weekly meetings where the progress of each part was confirmed. The progress was quite steady during the project but with a slight increase towards the end.

A.1.5 The amount of work

The overall amount of work was suitable for the scope of the course. We were able to divide the work evenly between ourselves and no one felt that there had been too much work. Due to the nature of our project most of the work was related to planning, conceptualizing and research while the implementation requires less work.

A.2 In what regard was the project successful?

We were able to provide Inclus with three possible solutions for the three-point estimate problem. In this sense, we were able to achieve the originally planned goal of this part of the project. We also provide the code used in this project to Inclus, so they can start experimentation right away.

We conceptualized a new approach for Inclus to elicit risk dependencies which can be seen even as a novel approach in the industry. We expect our approach to fulfill the goal Inclus presented well, but at the time of writing we had not had a meeting with Inclus and the client's reactions are as of yet unknown.

A.3 In what regard was project less successful?

The implementation was only on a conceptual level, and that is why the methods themselves were not the most creative or revolutionizing. The concepts in this topic might have had another level of exploration, but we kept it simple (which is not necessarily a bad thing in this context).

A.4 What could have been done better?

Scheduling could have had more effort from each team member, the completion of the project weighted on the later stretch of the spring. We could have ensured the scope much earlier, kept up tight communication between parties and reacted to feedback much quicker. The project might also seem a little unstructured and we could have concentrated on conjoining the different parts of the project.

A.5 Team

Weekly meetings were very important to our team as we had a member who was living in a different part of the country. Communication with inside the team was successfully conducted throughout the project. We also feel that no individual member had too much work compared to any other member.

A.6 Teaching staff

The teaching staff provided adequate support throughout the course. The teaching staff responded quickly to queries and agreed to host meetings when needed. The course was organized well and the structure of it seemed logical. All the meetings were smooth and their schedules were published in time. A bonus was also the ability to join remotely to the meetings.

A minor improvement could be made to the feedback given for the project plan and interim report. Our team had occasionally a hard time figuring out the meaning of some parts of the feedback because we couldn't understand the handwriting. One of these parts is presented in Figure 20.

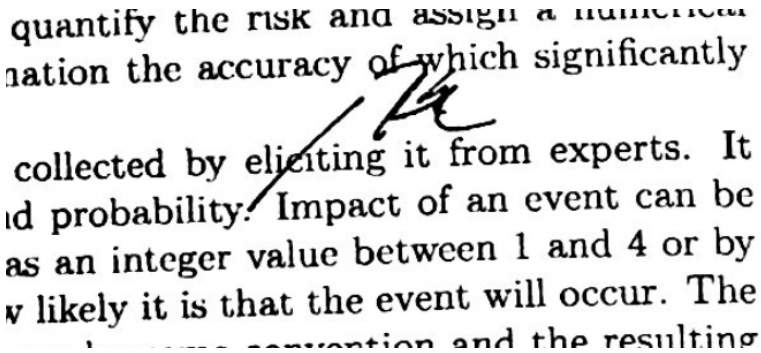

Figure 20: A confusing part of handwritten feedback.

A.7 Client organization

Inclus was very flexible and easily available even after office hours, which was very much appreciated by our team. Communication and working with Inclus proved to have no problems. Inclus gave us a lot of freedom to work on the project as we saw most fitting. This made our workflow easy, but at the same time it constituted some lack of active direction from the client's part. In our project their chosen approach of giving us plenty of freedom worked relatively well.