CIRRELT
1971 – 2021

CIRRELT-2021-51

A Machine Learning Approach to deal with Ambiguity in the Humanitarian Decision Making

Emilia Grass
Janosch Ortmann
Burcu Balcik
Walter Rei

December 2021

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

A Machine Learning Approach to deal with Ambiguity in the Humanitarian Decision Making

Emilia Grass[1], Janosch Ortmann[2], Burcu Balcik[3], Walter Rei[2,4,*]

1. Imperial College London, Institute of Global Health Innovation, London, United Kingdom
2. Department of Management and Technology, Université du Québec à Montréal, Montréal, QC, Canada
3. Ozyegin University, Industrial Engineering Department, Istanbul, Turkey
4. Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

Abstract. One of the major challenges for humanitarian organizations when planning relief efforts is dealing with the inherent ambiguity and uncertainty in disaster situations. The available information that comes from different sources in post-disaster settings may involve missing element sand inconsistencies, which can severely hamper effective humanitarian decision making. In this paper, we propose a new methodological framework based on graph clustering and stochastic optimization to support humanitarian decision makers in analyzing the implications of divergent estimates from multiple data sources on final decisions and efficiently integrating these estimates into decision making. We illustrate the proposed approach on a case study that focuses on locating shelters to serve internally displaced people in a conflict setting, specifically, the Syrian civil war. We use the needs assessment data from two different reliable sources to estimate the shelter needs in Idleb, a district of Syria. The analysis of data provided by two assessment sources has indicated a high degree of ambiguity due to inconsistent estimates. We apply the proposed methodology to integrate divergent estimates into the decision making for determining shelter locations in the district. The results highlight that our methodology leads to higher satisfaction of demand for shelters than other approaches such as a classical stochastic programming model. Moreover, we show that our solution integrates information coming from both sources more efficiently thereby hedging against the ambiguity more effectively.

Keywords: humanitarian decision making, ambiguity, data aggregation, clustering, Syrian conflict, needs assessment

---

* Corresponding author: rei.walter@uqam.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2021

1 Introduction

While the availability of high-quality information is crucial to make effective decisions for all organizations, it can be difficult to access complete and accurate information in some settings. In particular, the nature of the information flow in complex humanitarian environments (such as after the occurrence of a natural disaster or during a conflict) can significantly impede effective decision making processes of humanitarian agencies, whose mission is to provide timely and sufficient aid to the affected communities (Day et al., 2012; Altay, Labonte, 2014; Comes et al., 2020). Specifically, humanitarian agencies have to make decisions under significant uncertainty due to lack of sufficient information on various parameters (e.g., needs, vulnerabilities, infrastructure network conditions) that are critical for disaster response planning. Moreover, to estimate these parameters, agencies often need to make sense of a large amount of information with missing and inconsistent elements, which can create high degrees of ambiguity in decision making. Specifically, ambiguity is defined as the "uncertainty about probability, created by missing information that is relevant and could be known" (Snow, 2010). While it may not be possible to eliminate ambiguity in post-disaster environments, we propose a methodological framework that enhances agencies' capabilities to deal with ambiguity in decision making.

In post-disaster environments, available information may involve inconsistencies since data can come from a variety of sources (Day et al., 2012; Altay, Labonte, 2014). For instance, post-disaster needs may be estimated by using pre-disaster information about the affected region (e.g., governmental statistics) and post-disaster information obtained through various technologies (e.g., satellite pictures, aerial images collected by drones), media reports as well as interviews made by local key informants (such as community leaders, affected people, local agencies). In addition to the large number and diversity of information sources, different methods and assumptions can be used in data processing, which can lead to different estimates on critical parameters that are used for planning response activities. While considering all available information may be attractive in making plans, it is challenging for humanitarian organizations to systematically integrate different estimates into decision making in an environment where the pressure and stakes for acting quickly are high. There is an overarching need for approaches that support humanitarian decision-makers to integrate information processing and decision making in post-disaster settings effectively (Raymond, Al Achkar, 2016; Benini et al., 2017; O'Brien, 2017; Comes et al., 2020). In this study, we aim to address this important research gap.

When faced with multiple estimates on a parameter (e.g., affected population in a town, the proportion of people with shelter or food needs), a humanitarian decision-maker can combine different values into a single value by applying simple aggregation techniques such as taking the highest data value to "play it safe" (Day et al., 2012), or computing average or weighted-average values (Benini et al., 2017). Defining a triangular distribution based on the best, minimum and maximum estimates is also possible (Benini et al., 2017). In the humanitarian logistics literature, it is common to define probability distributions to represent the uncertainties brought by different estimates and then use stochastic optimization techniques to support post-disaster decisions (such as last mile relief distribution, evacuation planning, shelter location) (Liberatore et al., 2013; Dönmez et al., 2021). However, one disadvantage of such mathematical aggregation of data without examining its consequences on decision making is that it can mask the effects and contributions of individual data sources in final decisions (Benini et al., 2017). That is, when the data that comes from different sources is aggregated into a single value or a probability distribution in advance, it is not possible to observe whether the final solution would correspond to a consensus decision if the individual assessments

were considered. Therefore, one cannot identify which decisions are well supported by different estimates, and which ones are significantly affected by the differences among assessments. Moreover, decision-makers may not know which data aggregation techniques to use (such as computing simple averages or using more sophisticated techniques), and most importantly, what effects the chosen aggregation techniques will have on the final decisions. Therefore, additional information that would reduce such high level of ambiguity in decision making would be valuable (Snow, 2010).

In this study, rather than merging data that comes from different sources by using an aggregation method in advance of solving a decision-making problem, we aim to develop a method that can effectively integrate the data aggregation and decision making processes. Specifically, given different estimates provided by multiple data sources on critical parameters for post-disaster decision making, we present an approach based on stochastic optimization and unsupervised machine learning, specifically graph clustering. The aim of our graph clustering approach is to identify groups of scenarios whose associated solutions are similar. The resulting clusters provide the information that directly reduces the level of ambiguity faced by the decision-maker. More specifically, the proposed methodological framework aims to deal with ambiguity in humanitarian decision making by i) analyzing solutions systematically to identify whether there exists a high degree of consensus among different estimates in terms of their implications on decisions and observe how different estimates influence the decisions, and ii) integrating the data from different sources into decision making in a meaningful way by adjusting the weights to different solutions to obtain the most "agreed" solution. To the best of our knowledge, this is the first study to address ambiguity and the integration of divergent estimates into complex humanitarian decision-making processes.

While our methodology is general and can be applied to different decision-making environments where quantitative estimates are available from multiple sources, we illustrate the implementation of the proposed approach on a case study focusing on the integration of needs assessment data with shelter location decisions during the Syrian conflict. Since the beginning of the conflict, sector-specific (e.g., shelter, nutrition) needs across the country have been systematically assessed by different humanitarian initiatives. However, discrepancies may occur between different assessments since different assessment agencies may follow different methodologies to conduct surveys with different key informants, as well as they may use different assumptions and techniques while cleaning and aggregating the collected information. For instance, as reported by Benini et al. (2017), the estimated proportion of internally displaced people (IDP) in a single sub-district of Syria varies between 15% and 74% across different data sources. We apply the proposed methodology to the needs assessment data provided by two reliable assessment initiatives, which were collected in July/August 2018 from the Idleb sub-district of Syria. We integrate this needs assessment data related to the shelter needs of the affected population into decision making for designing a shelter network and show the benefits of the proposed approach in dealing with information ambiguity compared to traditional approaches.

The rest of this paper is organized as follows. In Section 2, we review the relevant literature. In Section 3, we define our problem, and in Section 4, we describe our methodological framework. We present a numerical analysis to illustrate the implementation and advantages of the proposed methodology in Section 5. Finally, we conclude and discuss future research in Section 6.

2 Literature Review

In this section, we review the relevant literature on information management in humanitarian operations (Section 2.1), scenario clustering methods (Section 2.2) and shelter location problems (Section 2.3).

2.1 Information Management in Humanitarian Operations

This study is motivated by the need for systematical approaches to facilitate linking information processing and decision making stages, which is a primary challenge in humanitarian environments. In all its benefits, growing amounts of data from heterogeneous sources can bring significant challenges for humanitarian organizations that have limited time to make decisions. The importance of accessing accurate information for effective humanitarian decision making and the difficulties of information management in disaster contexts have been widely discussed in the literature (e.g., Day et al. (2012); Altay, Labonte (2014); Walle Van de, Comes (2015); Gupta et al. (2016); Benini et al. (2017); Gupta et al. (2019); Comes et al. (2020)). Altay, Pal (2014) highlight that the quality of information is crucial for effective use of resources. Especially in the post-disaster phase, inaccurate and noisy information often leads to ambiguities that significantly hamper decision-making. Yin, Jing (2014) analyze the functioning of cognitive schemata, including information ambiguity, in the perception of disaster situations. As highlighted by Taylor et al. (2021), ambiguity and uncertainty are the main reasons for the difference between post-disaster policy formulation and its actual implementation. They present a framework that assess the impact of ambiguity and uncertainty as obstacles to policy implementation. According to Hosseinnezhad, Saidi-mehrabad (2018), information is often merged from several heterogeneous sources in disaster chains, where decision-makers might face contradictory information. Therefore, the authors emphasize the need for new approaches integrating ambiguity, vagueness and inconsistency. We aim to address the need for innovative methods to better link information management and decision making in humanitarian supply chains, which is increasingly stressed as an important research gap (e.g., Van Wassenhove, Besiou (2013); Comes et al. (2020)).

In this study, we address this challenge by implementing the proposed methodology based on unsupervised machine learning. The increasing use of technology in disaster settings enables the accessibility to ever greater amounts and types of data, making machine learning techniques increasingly popular in disaster management (Sokat et al. (2016); Swaminathan (2018)). For instance, Ofli et al. (2016) propose a parameterized classification model to identify damaged shelters or buildings based on aerial imagery. Different machine learning methods, e.g. Naïve Bayes, random forests, neural networks, are applied to data from social networks like Twitter to extract and categorize useful information in disaster situations (Li et al., 2018; Reynard, Shirgaokar, 2019; Dong et al., 2021). A broader overview of recent machine learning approaches in disaster and pandemic management can be found in Chamola et al. (2020). Our study also contributes to this stream of literature by introducing an application of machine learning techniques to humanitarian setting.

2.2 Scenario Clustering

As shown by several review papers (e.g., Grass, Fischer (2016b); Gutjahr, Nolz (2016); Yáñez-Sandivari et al. (2020)), discrete scenarios are most often used to capture the uncertainties in disaster contexts. In general, there are two ways of generating scenarios in a humanitarian setting, either by deriving them from past data on disasters or by interviewing experts (Yáñez-Sandivari et al., 2020). For instance, Andres et al.

(2020) propose a scenario-based artificial intelligence approach where scenarios are based on empirical data to forecast the number of forcibly displaced people.

In this study, we propose a scenario clustering approach to specifically analyze the levels of ambiguity regarding the source-specific scenarios. Scenario clustering techniques have been primarily used to search for patterns in, or associated with, scenarios or to reduce the number of scenarios. The generally large size of the scenario set (Birge, Louveaux, 2011) can lead to formulations that are intractable to solve directly (e.g., Dyer, Stougie (2006)). Decision-makers are faced with a trade-off between a sufficiently large set of meaningful scenarios on the one hand and the size of the scenario set on the other. This motivates the *scenario reduction problem*, namely to identify a subset that minimizes some approximation error induced by replacing the original set with the identified subset.

The *clustering approach* to scenario reduction aims to create a partition $C_1, ..., C_M$ of the scenario set $\mathcal{S}$. In other words, each scenario $s \in \mathcal{S}$ is contained in exactly one of the clusters $C_1, ..., C_M$. In general, the clusters are allowed to be of different sizes and chosen such that all elements of a cluster are similar to each other, subject to a given notion of similarity. The number of clusters M is an input parameter to be specified. Scenario reduction can then be performed, for example, by choosing one representative from each cluster. In this way, groups of similar scenarios are replaced by a single representative. See for example Jain, Dubes (1988) or Han et al. (2011) for an overview of clustering methods.

While there are many possible notions of similarity on which the clustering can be based, two have proven particularly useful in scenario reduction. First, *parameter based clustering* is based on the values of the uncertain parameters that the scenarios represent (Crainic et al., 2014). More concretely, suppose that the uncertain parameters are all real-valued, say $\alpha_1, ..., \alpha_d$. Each scenario represents particular values that these parameters can take. In this way, we can associate a d-dimensional vector to each scenario. The notion of distance is then chosen to be a distance on the space of d-dimensional vectors, for example the Euclidean (sum-of-squares) distance. In other words, two scenarios are considered to be close to each other if they represent similar values for the uncertain parameters. Since scenarios are embedded into a Euclidean space, standard clustering algorithms such as the k-means (Lloyd, 1957) or the k-medoids algorithm (Jain, Dubes, 1988) can be applied. This idea has been successfully applied, for example, in Crainic et al. (2014) to solve a stochastic network design problem.

The second, *solution-based approach* seeks to identify groups of scenarios whose associated solutions are similar. One example of this approach is given by Keutchayan et al. (2021). In Hewitt et al. (2021), an *opportunity cost distance* on scenarios is introduced. Under this distance, scenarios are considered close if they have mutually acceptable decisions associated to them. This gives rise to a weighted graph structure on the set of scenarios. Due to the more complex structure, the k-means type clustering algorithms need to be replaced by graph clustering methods (Shi, Malik, 2000; Luxburg von, 2007). Our approach uses and extends the methodology of Hewitt et al. (2021) by analyzing the level of decision agreement among scenarios and integrating these scenarios through optimization to reach a consensus decision.

2.3 Needs Assessment and Shelter Location Problems

In this study, we propose an integrated data aggregation and decision making methodology, which is illustrated on a post-disaster setting that focuses on linking the needs assessment data and shelter location decisions during a complex emergency. Both post-disaster needs assessment planning and shelter location

problems are widely studied in different humanitarian contexts (e.g., see the reviews by Galindo, Batta (2013); Farahani et al. (2020)). The needs assessment process focuses on collecting information from the affected communities to understand their needs for survival and well being (e.g., Balcik (2017)). While the assessment information may highly affect the design and management of relief operations, the linkages between assessment and response phases have not been explored yet. That is, in existing studies, data analysis and decision making are usually not considered in an integrated way; rather, available assessment data is processed first to estimate the values of uncertain critical parameters (i.e., demand), which are then used as deterministic or stochastic inputs to solve an optimization problem for making disaster response decisions (e.g., Stauffer et al. (2016); Lorca et al. (2017)). In contrast to the traditional sequential approach, we present a new method that integrates the available needs assessment data into decision making for disaster response, which can provide more intuition to decision-makers in understanding the effects of data aggregation and making sense of different solutions generated by data from different assessment sources.

Locating shelters such as town halls, gyms or tents, to serve the affected people after a disaster is an active research field (Kılcı et al., 2015; Jahre et al., 2016; Ni et al., 2018; Kınay et al., 2018; Azizi et al., 2021). Given that location decisions are extremely impeded by the high degree of uncertainty inherent in disaster and crises situations, stochastic optimization techniques are widely utilized (Liberatore et al., 2013; Dönmez et al., 2021). Specifically, two-stage stochastic models have been often used to model uncertainty (e.g., Grass, Fischer (2016a); Elçi, Noyan (2018); Paul, Zhang (2019)). These models consist of decisions made before (i.e., first stage) and after (i.e., second stage) the realization of uncertainty represented by scenarios. Two-stage stochastic programming is well suited in the chaotic aftermath of a disaster where there exist a high level of uncertainty regarding needs in the affected region. We consider a two-stage stochastic model to locate shelters with limited capacities in the aftermath of a disaster, and explore how ambiguous needs assessment information can be integrated in the decision making. Note that robust optimization, particular distributionally robust optimization, is an approach that can be applied to solve problems that involve ambiguity and to find solutions that hedge against the risks associated with this ambiguity. This is done by considering the worst case across the ambiguity, see for example the review by Rahimian, Mehrotra (2019). However, our objective here is to allow the decision-maker to analyze and link the decisions to be made with the information provided by the different data-sources, which cannot be achieved by applying robust optimization.

This study contributes to the literature by developing a new methodology that links information processing with decision making in a post-disaster environment that involves uncertainty as well as ambiguity and presenting the benefits of the proposed approach on a complex emergency setting with real data. The proposed methodology can support humanitarian decision-makers to eliminate the excessive effort and energy spent to deal with information ambiguity without connecting it to decisions, and hence shifting the focus from aggregation of data to aggregation of data with respect to conclusions to be drawn, thereby allowing humanitarian organizations to obtain solutions supported by different viable assessments. Although the proposed approach is illustrated on a shelter location problem formulated as a two-stage stochastic model, it is general and would apply to any kind of optimization model involving scenarios.

3 Problem Definition

In this section, we first define the problem in general terms (Section 3.1) and then introduce a shelter location problem in a humanitarian setting (Section 3.2).

3.1 General Problem Statement

Consider a decision maker who faces a given problem involving uncertainty, such as the allocation of relief resources under demand or supply uncertainty. Specifically, the decision maker must make a series of decisions, which we represent as the variable vector x, while the informational context in which the problem appears contains uncertain parameters, which we represent as the parameter vector ξ. We further assume that $\phi(x, \xi)$ defines the function that the decision maker seeks to optimize. Without loss of generality, let us assume that function $\phi(x, \xi)$ computes the total value associated with x if the uncertain parameters turn out to be ξ and which the decision maker is interested in maximizing. Considering that vector ξ contains a series of uncertain parameters, then for a fixed set of decisions x, $\phi(x, \xi)$ defines a distribution of values (i.e., each one associated with a possible realization of vector ξ).

In the context of our shelter location problem (Section 3.2), x is the choice of shelter locations to serve the affected population that need shelter, whereas ξ represents a number of uncertain parameters that affect the outcome of the allocation of aid, such as the number of people in need of shelter. The function $\phi(x, \xi)$ then represents the total number of IDPs that can be accommodated if a decision x is taken and the realization of the uncertain parameters is ξ.

The probability measure $\mathbb{P}$ encodes the distribution of the vector of uncertain parameters ξ. The following optimization model can then be solved by the decision maker to find an appropriate solution to the problem:

$$\max_{x \in A} \mathbb{E}\left[\phi(x, \xi)\right], \tag{1}$$

where A defines a set of constraints that are imposed on the decision variables x. The objective function defined in model (1) is the expected value of a given solution and it represents what is often referred to as the value function or recourse function in a stochastic program (Birge, Louveaux, 2011). We seek to maximize the total expected number of people that can be accommodated in shelters. It is assumed that a series of data sources, which are different assessments for shelter needs, are leveraged to formulate the probability measure $\mathbb{P}$. Let K define the finite set of distinct data sources that are considered. It is further assumed that each data source $k \in K$ can be used to define a source-specific probability measure, which we define as $\mathbb{P}^k$. Moreover, the applied hypothesis is that the same level of confidence is associated with all the source-specific probability measures $\mathbb{P}^k$, $\forall k \in K$. Therefore, there is ambiguity regarding which of the probability measures should be used to define model (1).

Stochastic optimization enables problems to be solved by formulating the uncertain parameters using a probability measure that is explicitly defined, see Birge, Louveaux (2011). Although this approach does not directly tackle ambiguity, it allows a problem to be solved using different probability measures. When the approach is applied to the present problem, given any $\mathbb{P}$, a set S of scenarios with associated probabilities p_s for $s \in \mathcal{S}$ is generated to produce a more manageable problem to solve. Thus, the following discrete probability measure is obtained:

$$P_{\mathcal{S}} = \sum_{s \in \mathcal{S}} p_s \delta_s \tag{2}$$

where δ_s, $\forall s \in \mathcal{S}$, define indicator functions that state whether or not the associated scenarios appear in a given random experiment. Another way of viewing (2) is as a discretization of $\mathbb{P}$. Assuming that ξ_s represents the realization of the uncertain parameters associated with scenario $s \in \mathcal{S}$, then the following approximation problem (i.e., with respect to the original problem (1)) can be solved:

$$\max_{x \in A} \sum_{s \in \mathcal{S}} p_s \phi(x, \xi_s). \tag{3}$$

Assuming that problem (3) is solved using a given set $\mathcal{S}^k$, that is generated using the source-specific probability measure $\mathbb{P}^k$, then one would obtain the optimal solution $x_k^\star$. Specifically, solution $x_k^\star$ defines a set of feasible decisions (i.e., $x_k^\star \in A$) that provide the maximum approximated value function if the data source $k \in K$ is used to generate the scenario set $\mathcal{S}^k$ (i.e., the underlying assumption being that $\mathbb{P}^k$ defines the distributions of the parameters ξ). If this two-step process $\big[$*Step 1*: generate a set of scenarios; *Step 2*: solve the resulting approximated problem (3)$\big]$, is then repeated for all available data sources k, then one obtains a set of feasible (and most likely different) solutions $x_k^\star \in A$, $\forall k \in K$. Each of these solutions prescribes the set of decisions that would be appropriate to implement if each data source is used separately to formulate the probability measure applicable to formulate the distributions of the uncertain parameters. On their own, each solution $x_k^\star$ does not guarantee an efficient integration of the probabilistic information that may be gathered from the other data sources (i.e., $\forall k' \in K$ such that $k' \neq k$). Solution $x_k^\star$ only provides the perspective of what decisions are warranted if $\mathbb{P}^k$ is trusted to properly formulate the prevailing uncertainty. However, $x_k^\star$, $\forall k \in K$, can be used as the basis to evaluate just how close a given solution comes to simultaneously reaching the prescribed decisions when the probabilistic information, inferred from each data source, is considered. In particular, given a specific solution to the considered problem $x \in A$, let us define the following function:

$$\epsilon_k(x) = \sum_{s \in \mathcal{S}^k} p_s \phi(x_k^\star, \xi_s) - \sum_{s \in \mathcal{S}^k} p_s \phi(x, \xi_s). \tag{4}$$

Function $\epsilon_k(x)$ defines the gap, evaluated based on the approximated probabilistic model derived using the data source k, associated with solution x when it is compared with the optimal solution $x_k^\star$ (i.e., which is obtained under the assumption that $\mathbb{P}^k$ is applicable). An overall gap can then be defined as follows:

$$\epsilon(x) = \sum_{k \in K} \epsilon_k(x). \tag{5}$$

To deal with the ambiguity encoded in the probability measure, we then propose to search for a solution $x^\star$ that minimizes the overall gap value:

$$x^\star \approx \arg\min_{x \in A} \epsilon(x). \tag{6}$$

In the present paper, we will show that, by using a novel clustering methodology to perform a systematic analysis of the scenarios included in $\mathcal{S}^k$, $\forall k \in K$, we can define an alternative approximation model of type (3) that can be solved to obtain a high-quality solution of type (6).

3.2 Shelter Location Problem and Model

As stated in the introduction, when considering the type of problems that are faced by humanitarian organizations (such as the deployment of aid in post-disaster environments) another important imperative for

decision makers is the need to analyze how the various data sources $k \in K$ impact the decisions to be made (i.e., $x \in A$). From a qualitative perspective, it can be valuable for them to gain insights regarding how the various data sources influence their decisions. Such insights are often essential to justify the choices made regarding how the aid is deployed and the available resources managed. But such analysis may also be required for accountability purposes with respect to donors, who expect that the use of their donations be determined following a careful needs assessment. In all cases, properly integrating the information provided by the various data sources directly into the decision processes defines an important challenge in humanitarian planning settings.

In this subsection, we consider a problem of accommodating people or families affected by a disaster, e.g. a civil war as in our case, where it is difficult to obtain accurate information. For our case study in Section 5, we use two data sources, i.e. $|K| = 2$, that collect information to assess shelter needs in crisis-affected regions based on different surveys made in the same district in close periods. Our network consists of nodes, e.g. cities or districts, where shelter demand can arise and where facilities such as a tent or a public building can be set up or temporarily converted to meet this demand. In the first stage, i.e., before the full extent of the disaster and the demand have been realized, decisions on shelter locations have to be taken. Each shelter can accommodate people within a particular coverage distance. When the actual number of people and families in need of shelter is known, second-stage decisions on how many of them can be accommodated are taken. The objective of our model is to meet the expected demand for sheltering where the number and capacity of shelters are limited. The stochastic optimization model presented in the following is an adopted and simplified version of the one proposed by Noyan et al. (2015). We present our notation and the mathematical model below.

Sets:

I: set of demand nodes

O: set of candidate shelter nodes

$M_o = \{i \in I | D_{io} \leq \tau\}$: set of demand nodes that can be covered by shelter at node $o \in O$

$N_i = \{o \in O | D_{io} \leq \tau\}$: set of candidate shelters that can cover demand at node $i \in I$

S: set of possible scenarios

Scenario-independent parameters:

D_{io}: distance between demand node $i \in I$ and shelter at node $o \in O$

G: maximum available shelter capacity, i.e. total number of people that can be accommodated

κ: maximum number of shelters to be opened

h_o: maximum number of people that can be accommodated in shelter $o \in O$

τ: maximum coverage distance

Scenario-dependent parameters:

q_i^s: number of people in need of shelter at node $i \in I$ in scenario $s \in S$

p_s: probability of scenario $s \in$ S

$\theta^s = min\left\{G, \sum_{i\in I} q_i^s\right\}$: minimum value of available shelter capacity and the total number of people in need of shelter in scenario $s \in S$

Scenario-independent decision variables (first stage):

z_o: 1 if a shelter is opened at node $o \in O$; 0 otherwise

Scenario-dependent decision variables (second stage):

r_o^s: number of people accommodated in shelter at node $o \in O$ in scenario $s \in S$

y_{io}^s: 1 if demand at node $i \in I$ is covered by shelter at node $o \in N_i$ in scenario $s \in S$, 0 otherwise

Formulation:

$$\max \quad \sum_{s\in S}\sum_{o\in O} p_s r_o^s \tag{7}$$

$$\text{s.t.} \quad \sum_{o\in O} z_o \leq \kappa \tag{8}$$

$$\sum_{o\in O} r_o^s \leq \theta^s \quad \forall s \in S \tag{9}$$

$$\sum_{o\in N_i} y_{io}^s \leq 1 \quad \forall i \in I,\, s \in S \tag{10}$$

$$r_o^s \leq h_o z_o \quad \forall o \in O, s \in S \tag{11}$$

$$r_o^s \leq \sum_{i\in M_o} q_i^s y_{io}^s \quad \forall o \in N_i, s \in S \tag{12}$$

$$y_{io}^s \leq z_o \quad \forall i \in I,\, o \in N_i,\, s \in S \tag{13}$$

$$z_o \in \{0,1\}, \quad \forall o \in O \tag{14}$$

$$y_{io}^s \in \{0,1\} \quad \forall i \in I,\, o \in N_i,\, s \in S \tag{15}$$

$$r_o^s \geq 0, \quad \forall o \in O,\, s \in S. \tag{16}$$

The objective in (7) is to maximize the expected number of accommodated people. This has to be achieved by meeting the following constraints. No more than κ shelters can be opened in the first stage, imposed by constraint (8). Second-stage constraints (9) state that in each scenario s the capacity of shelter at node o cannot exceed θ^s, representing the minimum between available shelter capacity G and the overall number of people in need of shelter. People at node i can be accommodated by at most one shelter that is located within distance τ, i.e., within a certain coverage distance, which is expressed by constraints (10). Constraints (11) and (12) ensure that shelters cannot accommodate more people than there is capacity and shelter demand, respectively. According to constraints (13), only those facilities can provide shelter that are open and located within coverage distance τ. Binary variables z_o, y_{io}^s and non-negative variables r_o^s are defined in (14)-(16).

4 Methodological Framework

We now detail the proposed methodological framework, which enables a large amount of information contained in the assessments emanating from the set of data sources $k \in K$ to be more efficiently integrated within decision making. These data sources can be used to specify a probability estimate for an event or a state,

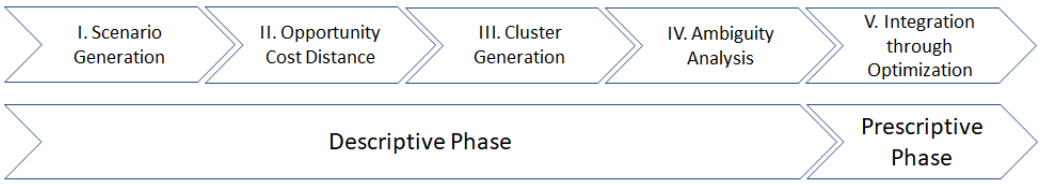


Figure 1: General methodological process.

or simply to provide a range of values (i.e., the minimum, maximum and most probable) for an unknown quantity such as the number of people in need, see Benini et al. (2017). In the latter case, the range of values can be used to define probabilistic measures, for example via the use of triangular distributions, which are easy to understand and interpret (Benini et al., 2017).

As discussed in the previous section, since the different data sources $k \in K$ may lead to drastically different assessments of the uncertain parameters, integrating the overall contextual information that is provided (i.e., the value vectors ξ_s, $\forall s \in S^k$ and $\forall k \in K$) becomes quite challenging for humanitarian organizations. To efficiently incorporate the ambiguous information provided by the set of data sources $k \in K$ to find a high-quality solution of type (6), we propose a two-phase methodological framework, as illustrated in Figure 1.

In the first phase (**descriptive phase**), a descriptive analysis is performed on the source-specific probability measures obtained from the set of data sources. The objective of this phase is not only to specify the information provided by the data sources, but also to assess the impacts that this information has on the considered planning problem. Upon completing this phase, knowledge is obtained on both the unknown contextual information of the problem and on the level of overall decision agreement that may exist between the models generated from the data sources with regards to how their information affects the problem.

The second phase of our framework is dedicated to the use of the obtained knowledge to prescribe an appropriate solution to the problem (**prescriptive phase**). Through the use of novel decision analysis techniques and mathematical programming methods, the information extracted from the data sources is efficiently interpreted and aggregated to provide decision support. Specifically, we will show how an alternative approximation model of type (3) can be defined to obtain a *consensus solution* $x^\star$ as defined by (6).

In the rest of the section, we describe the two phases included in the framework. The descriptive phase is explained in Section 4.1, while the prescriptive phase is presented in Section 4.2.

4.1 Descriptive Phase

As indicated in Figure 1, the descriptive phase consists of performing the following four distinct steps: scenario generation, the computation of the opportunity cost distance, cluster generation and ambiguity analysis.

Step I: Scenario generation. Obtaining information from each data source is subject to two types of error (Hoffman, Hammonds, 1994). On the one hand, there is the uncertainty encoded in the data source which we call *intrinsic uncertainty*. It is this type of uncertainty that motivates giving a range, rather than a point estimate. On the other hand, there is uncertainty not encoded in the data source, or *extrinsic uncertainty*. For example, any data source expressed through an expert assessment is likely subject to overestimation of the precision regarding the expert's predictions (Hammitt, Shlyakhter, 2006). Also, unlikely outcomes may not have occurred (or be explicitly considered) in the data set, which leads to their probability being underestimated (Abdellaoui et al., 2011). In the extreme case, the range of values for an

uncertain parameter obtained from different data sources may not even overlap: all values that lie in the possible range extracted from one data source may be considered impossible by the other.

In order to hedge the risk posed by this extrinsic uncertainty, we formulate a larger prediction uncertainty than that given by any individual data source (see Section 5.2 for more details). Let us recall that we denote by $\mathbb{P}^k$ the source-specific probability distribution associated to data source $k \in K$. That is, $\mathbb{P}^k$ encodes the assessment of uncertainty represented by the data source k. In our case study, we consider two data sources, which provide needs assessment results based on different surveys made in the same district in close periods. Recall further that x denotes the decision vector – in our case the allocation of shelter nodes – and that ξ denotes the vector of uncertain parameters, i.e., the shelter needs.

From these probability distributions, we then sample discrete values for the uncertain parameters and include them into scenarios: each scenario being associated with one set of values that the uncertain parameter vector takes. See King, Wallace (2012) for more details on sampling methods that can be applied in this context. In the following, we will denote the discretization of the probability measure $\mathbb{P}^k$ by $\mathcal{S}^k$ (the *scenario set*). For each scenario $s \in \mathcal{S}^k$ we denote by ξ_s the corresponding realization of the uncertain parameter ξ. Denoting by N_k the number of scenarios contained in $\mathcal{S}^k$ we can write

$$\mathcal{S}^k = \left\{s_1^k, \ldots, s_{N_k}^k\right\} \text{ and } \Xi^k = \left\{\xi_{s_1^k}, \ldots, \xi_{s_{N_k}^k}\right\}.$$

The sets containing all scenarios and their associated realizations of the uncertain parameters are denoted by

$$\mathcal{S} = \bigcup_{k \in K} \mathcal{S}^k \text{ and } \Xi = \bigcup_{k \in K} \Xi^k.$$

We assume throughout that the scenario sets generated from each data source are disjoint. Therefore,

$$|\mathcal{S}| = \sum_{k \in K} \left|\mathcal{S}^k\right| = \sum_{k \in K} N_k.$$

Step II: Opportunity cost distance. The second step of the descriptive phase defines the basis over which the scenarios included in the sets $\mathcal{S}^k$, $\forall k \in K$, will be compared and analyzed. Specifically, the idea is to interpret the information contained in ξ_s, $\forall s \in \mathcal{S}$, in terms of the decisions to be made regarding the specific decision making problem that is considered. Therefore, for each data source $k \in K$, the following solutions are obtained:

$$x(s_i^k) = \arg\max_{x \in A} \phi(x, \xi_{s_i^k}), \quad i = 1, \ldots, N_k. \tag{17}$$

These solutions can be understood as follows: if one is somehow able to predict that scenario s_i^k will occur (i.e., the data source k has thus provided the correct assessment) then the solution which should be chosen and implemented is $x(s_i^k)$, which is obtained by solving the problem (17) using the predicted scenario s_i^k. Each data source $k \in K$ is thus associated with the following solution set:

$$X^k = \left\{x(s_1^k), \ldots, x(s_{N_k}^k)\right\}.$$

The overall set of all such solutions is thus denoted as:

$$X = \bigcup_{k \in K} X^k.$$

Therefore, for $s \in \mathcal{S}$, we define $x(s) \in X$ as the representative solution that is associated with the considered scenario.

We now apply a notion of distance between scenarios, called *opportunity cost distance* that was first introduced in Hewitt et al. (2021). For any pair of scenarios $s_1 \in \mathcal{S}$ and $s_2 \in \mathcal{S}$, we evaluate the cost of predicting scenario s_1 and taking the associated decision, when in fact scenario s_2 occurs. Thus, these two scenarios are close with respect to this distance if the decisions associated to them are mutually acceptable (i.e., solutions $x(s_1)$ and $x(s_2)$ are good surrogates for one another). Mathematically, the opportunity cost distance is given by

$$d\left(s_1, s_2\right) = \phi(x\left(s_1\right), \xi_{s_2}) - \phi\left(x\left(s_2\right), \xi_{s_2}\right) + \phi(x\left(s_2\right), \xi_{s_1}) - \phi\left(x\left(s_1\right), \xi_{s_1}\right). \tag{18}$$

An opportunity cost distance matrix is then obtained by calculating the distance values using equation (18) for all scenario pairs in the overall set (i.e., compute $d\left(s_1, s_2\right)$, $\forall s_1, s_2 \in \mathcal{S}$).

Step III: Cluster generation. Equipped with the opportunity cost distance function, and having computed the associated distance matrix, we now look for groups of scenarios that are very close to each other, but relatively far away from the other groups. This step reduces to solving a clustering problem over the scenario set $\mathcal{S}$, for which various unsupervised machine learning methods can be applied, e.g., Shi, Malik (2000) and Luxburg von (2007). In the present case, we choose the normalized N-Cut algorithm (Shi, Malik, 2000; Hewitt et al., 2021), which seeks to minimize the diameter of each cluster in relation to the distance between clusters. In this way, we obtain a partition $C_1, \ldots, C_M$ of the scenario set $\mathcal{S}$ such that elements of the same cluster C_j are relatively close with respect to the opportunity cost distance (18), whereas members of two different clusters C_i and C_j for $i \neq j$ are relatively far away from each other. The number of clusters M can be chosen by the user depending on the context by considering the trade-off between a higher quality of the clustering (more clusters) and lower computational complexity (fewer clusters). In some contexts, M may be set in advance.

In our case, we will choose M so as to maximize a particular notion of clustering quality called the *Silhouette score*. This score is a measure of how close each scenario is to other members of its own cluster, compared to its distance to other clusters (Rousseeuw, 1987).

Step IV: Ambiguity analysis. The descriptive phase ends with a step that is dedicated to the analysis of the obtained clusters with a focus on diagnosing the level of decision agreement among the scenarios included in $\mathcal{S}$ and data sources K. Therefore, we begin this step by identifying how (if at all) the data sources agree with each other in terms of the most appropriate decisions to be made, by analyzing the clusters generated above. For any subset $U \subseteq \mathcal{S}$ we can define the *decision level of agreement*: $\Delta(U) \in [0, 1]$, by

$$\Delta(U) = \frac{4}{|U|^2} \sum_{s_1, s_2 \in U} \Delta\left(x\left(s_1\right), x\left(s_2\right)\right), \tag{19}$$

where $\Delta(x_1, x_2)$ denotes the normalized Hamming distance between two permissible solutions $x_1, x_2 \in A$, which is defined as follows:

$$\Delta(x_1, x_2) = \frac{1}{L} \sum_{l=1}^{L} 1_{x_1(l) \neq x_2(l)},$$

where L is the common length of x_1 and x_2, that is $x_1, x_2 \in \mathbb{R}^L$.

In this way, we can calculate the decision level of agreement within the clusters, i.e., $\Delta(C_j)$ for $j = 1, \ldots, M$. In addition, by computing $\Delta(\mathcal{S}^k)$, we can also measure the variance of the information obtained from one data source $k \in K$, i.e., to what extent the different scenarios generated from k lead to the same solutions (or decisions).

For a concrete example, suppose that there are five potential shelter locations enumerated 1,...,5 and that there are two potential solutions x_1 and x_2. If solution x_1 proposes to open shelters at locations 1, 2 and 3, and solution x_2 proposes to open shelters at locations 2, 3 and 4, there is agreement on opening shelters in locations 2 and 3, and not to open a shelter in location 5. In other words, the two solutions disagree on the opening of a shelter in two locations (1 and 4), so that $\Delta(x_1, x_2) = \frac{2}{5} = 0.4$. The metric Δ provides a value between 0 and 1 that allows decision-makers to quickly compare agreement between solutions x_1 and x_2: in the extremes, $\Delta(x_1, x_2) = 0$ means that the two solutions are identical whereas $\Delta(x_1, x_2) = 1$ implies that the two solutions disagree about whether to open a shelter or not at every single location.

Another important dimension to consider in this analysis is the distribution of scenarios' *origin* within a cluster. We will be interested in distinguishing between clusters where all scenarios were generated by a single data source and clusters with a mix of scenarios from different data sources. In other words, we analyze the distribution of data sources in a cluster. By explicitly considering this information, the decision-maker is able to directly analyze the levels of ambiguity related to the overall assessments provided by the different data sources (i.e., the context information contained in Ξ). Therefore, the more data sources are present in a given cluster, the less ambiguity is involved between them regarding the scenarios contained within the cluster. In other words, even though the scenarios may originate from different data sources and may specify different values for the uncertain parameters, they all lead to make decisions (i.e., find solutions to the problem) that are similar (i.e., solutions that are good surrogates for one another). This analysis thus provides value for an ambiguity-averse decision-maker. Next, we show how a measure can be defined to quantify such observations. More precisely, for a cluster C_j and a data source $k \in K$, let $\pi_k(C_j)$ be the proportion of scenarios in C_j generated from the data source k:

$$\pi_k(C_j) = \frac{\left|C_j \cap \mathcal{S}^k\right|}{|C_j|}. \tag{20}$$

We say that a data source $k \in K$ is *present* in a cluster C_j if $\pi_k(C_j) > 0$. We then define the *diversity of data sources within a cluster* via the entropy

$$H(C_j) = -\sum_{k \in K} \pi_k(C_j) \log(\pi_k(C_j)), \tag{21}$$

with the usual convention that $0 \log(0) = 0$.

The value of $H(C_j)$ lies between 0 and $\log(|K|)$ (recall that $|K|$ is the number of data sources). A value close to 0 indicates a low diversity of data sources. The extreme case of $H(C_j) = 0$ means that all scenarios in C_j were generated by a single data source. While a large value of $H(C_j)$ indicates a high diversity of data sources. The highest possible value of $H(C_j)$, namely $\log(|K|)$, means that every data source is present in the cluster with the same proportion.

To illustrate this, consider the case where $|K| = 2$, i.e., there are two data sources, say source #1 and source #2. Suppose that cluster C_1 contains only scenarios generated by data source #2. Then the diversity of sources is $H(C_1) = 0$. If cluster C_2 contains 10 scenarios from source #1 and 30 scenarios from source #2

then the diversity of sources is

$$H(C_2) = -\left[\frac{1}{4}\log\left(\frac{1}{4}\right) + \frac{3}{4}\log\left(\frac{3}{4}\right)\right] \approx 0.56.$$

Finally, suppose that cluster C_3 contains 20 scenarios each from the two sources. As shown above, $H(C_3)$ then takes the maximal value of $\log(2) \approx 0.69$. This is illustrated in Figure 2.

(a) (b)

Figure 2: (a) Entropy for two data sources. The x-axis represents the proportion π_1 of scenarios from data source #1, while the y-axis shows the corresponding entropy H. The three values for π_1 (0, $\frac{1}{4}$ and $\frac{1}{2}$) are marked in red. (b) Illustration of the example from the text: $H(C_1) = 0$, $H(C_2) \approx 0.56$ and $H(C_3) = 0.69$.

In the following, we show how this measure can be directly leveraged to find a consensus solution for the problem that is considered.

4.2 Prescriptive Phase

As indicated in Figure 1, the prescriptive phase consists of performing the integration through optimization to achieve a consensus decision.

Step V: Integration through optimization. In order to integrate the different estimates coming from various sources, we introduce two choices, namely a subset $\overline{\mathcal{S}}$ of the scenario set and weights w_s for each scenario $s \in \overline{\mathcal{S}}$, based on the metrics defined above. As a way of formalizing the problem expressed in (6), we define a *consensus solution* as follows

$$x^\star = \arg\max_{x \in A} \sum_{s \in \overline{\mathcal{S}}} w_s \phi(x, \xi_s). \tag{22}$$

This raises questions when formulating problem (22): which scenarios should be included in $\overline{\mathcal{S}}$ and how should the weights w_s be defined?

Regarding the choice of $\overline{\mathcal{S}}$, we could include all scenarios: $\overline{\mathcal{S}} = \mathcal{S}$. Then the consensus solution is obtained by explicitly considering all information stemming from the data sources. This would minimize the risk of not taking into account some of the information contained in the data sources. However, the size of the overall scenario set $\mathcal{S}$ might be very large, and considering the complexity involved in computing the value function ϕ, solving problem (22) with the full set of scenarios might not be computationally efficient. In this case, a representative scenario can be identified for the cluster and used as a proxy for the cluster in the definition of

(22). As proposed in Hewitt et al. (2021), the *medoid* of the cluster (i.e., the scenario that has the minimum average dissimilarity to all other scenarios of the cluster) can serve as the representative. Applying such a reduction, i.e., choosing $\overline{\mathcal{S}} \subset \mathcal{S}$, naturally leads to an approximation error with respect to using the full set $\mathcal{S}$ when searching for a consensus solution (22). That being said, as numerically illustrated in Hewitt et al. (2021), the use of the *medoids* as representatives of the clusters can still be used to produce a high-quality upper bound that can be more efficiently computed.

We define the weight w_s associated with a given scenario $s \in \mathcal{S}$ in two parts: 1) through the diversity of data sources within the cluster to which s belongs, and 2) according to the stochasticity of the data source from which s was generated. If a scenario reduction approach is applied to obtain $\overline{\mathcal{S}}$, then the weights associated with the scenarios in a given cluster are assigned to its respective representative.

In the first part, we place more weights on scenarios in clusters that contain more data sources. This is done as a means to prioritize the context information emanating from a cluster where there is less ambiguity related to the data sources that are present within it. When the data sources provide a differing view on the underlying uncertainty, this can lead to a skewed representation of the information sources in clusters. Recall that in our setting we cannot judge the reliability of each source and each source is assigned the same level of confidence. Thus, a source whose information leads to a higher level of uncertainty in our model will be represented in a larger number of different clusters. In turn, this knowledge allows us to better hedge against the risks of inaccurate predictions. This motivates the second part, where we place more weight on scenarios generated by data sources that appear in more clusters.

Diversity weight. The first weight $w_j^{(1)}$ is the same for each scenario in a given cluster C_j, i.e., the weight only depends on the cluster index $j \in \{1, ..., M\}$ and defined as follows (recalling the definition of H from (21)):

$$w_j^{(1)} = \lambda_K + H\left(C_j\right), \qquad \text{where} \quad \lambda_K = \frac{\log\left(|K|\right)}{4}. \tag{23}$$

Stochasticity weight. As explained above, we also place more weight on scenarios generated from data sources that appear in more clusters. The second weight is the same for each scenario that was generated from the same source. We therefore denote the second weight by $w_k^{(2)}$ for $k \in K$ (recall that K is the set of data sources).

Suppose that two scenarios s_1 and s_2 were chosen uniformly from $\mathcal{S}^k$, the set of scenarios generated by source k. The weight $w_k^{(2)}$ is an affine function of the probability that s_1 and s_2 belong to different clusters. In other words, the weight is higher if the source k is more evenly represented across the clusters. More formally, let $\iota \colon \mathcal{S} \longrightarrow \{1, .., M\}$ denote the function that maps each scenario s to the index $\iota(s)$ of the cluster to which it belongs, i.e., so that $s \in C_{\iota(s)}$. Then

$$w_k^{(2)} = \frac{1}{4} + \frac{1}{\left|\mathcal{S}^k\right|^2} \sum_{j=1}^{M} \left|C_j \cap \mathcal{S}^k\right| \left|\mathcal{S}^k \setminus C_j\right|. \tag{24}$$

The addition of $\frac{1}{4}$ in the formula is important in the case where all scenarios generated by a data source lie in the same cluster.

Defining the overall weight. Having defined the two weights $w_j^{(1)}$ and $w_k^{(2)}$, we now define an overall weight on each scenario by multiplying them together. Recall that $w_j^{(1)}$ only depends on s through the cluster

j that s belongs to and $w_k^{(2)}$ only on the data source k scenario s was generated from. However, this is not quite satisfactory yet, since we would like the weights to be equal to 1 on average.

Formally, we define the overall weight w_s for $s \in \mathcal{S}$ as:

$$w_s = \frac{w_{\iota(s)}^{(1)} w_{\gamma(s)}^{(2)}}{W}, \qquad \text{where} \quad W = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} w_{\iota(s)}^{(1)} w_{\gamma(s)}^{(2)} \tag{25}$$

and $\gamma(s) \in K$ denotes the data source from which scenario s was generated, that is $s \in \mathcal{S}^{\gamma(s)}$. The definition of the normalization constant W ensures that the average of the weights is equal to 1:

$$\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} w_s = 1.$$

An illustrative example can be found in Appendix A.

5 Numerical Study

In this section, we present a numerical study developed based on data from the Syrian conflict to illustrate the implementation of the proposed methodology and assess its value for decision-makers. We focus on the integration of the needs assessment data with decision making for locating shelters to serve the people in need. We first give background information on the available needs assessment data provided by different sources (Section 5.1). We then explain the implementation of the proposed methodology to this setting (Section 5.2) and conclude with the corresponding results and analyses performed (Section 5.3).

5.1 Case Data Set

Syria has been at civil war since 2011, which has led to millions of casualties and displaced people (UN Refugee Agency, 2021). In the light of the hazardous circumstances in Syria, gathering accurate information on the humanitarian situation is extremely challenging. Various humanitarian initiatives conduct needs assessments in the affected regions to gather information on the community necessities. The collected information is processed (i.e., cleaned, combined, cross-checked with secondary sources) and the sector-specific needs (e.g., shelter, nutrition) in each district are published publicly.

We focus on two major assessment data sets, which are made publicly available by two humanitarian initiatives, namely the Humanitarian Needs Overview (HNO) and REACH. HNO (2019) provides estimates on the number of people in need for different types of relief in each district of Syria. We consider the nation-wide needs assessment of HNO conducted for 6,322 communities in Syria between July and August 2018. Specifically, 95,000 surveys at the household level were carried out. REACH (2018) also conducts need assessments in Syria on a regular basis since 2012. The assessments are based on community-level interviews by key informants, which are selected based on their knowledge of resident populations and IDPs in the community and sector-specific expertise. Specifically, three to seven key informants at each location are interviewed. In needs assessment reports, REACH provides the estimated total number of people residing in a district and the percentage of people requiring different types of supplies, e.g., water, medical items, food and shelters. We consider the assessment data set of REACH based on the interviews conducted between 12 and 20 August 2018. In the following, we refer to HNO as source #1 and REACH as source #2 who provides estimates on the humanitarian needs.

In both data sets, we focus on the assessments of Idleb district, which is located in northwestern part of the country bordering Turkey. Idleb is one of the most tormented parts of Syria due to frequent skirmishes between the Syrian government and the opposition forces. Due to the recurring bombardment and air strikes, about 1.7 million people have fled the area seeking security in neighboring countries like Turkey. Those who stay require essential supplies like water, food and medical care. Idleb district consists of 26 sub-districts, which are represented by nodes placed in the center of each sub-district.

To illustrate our approach, we chose one item type for simplification. Specifically, we focus on people in need of shelter in Idleb. While source #1 provides estimated number of people requiring shelter in detail, source #2 provides an aggregate estimate, which specifies that about 56% of local people are in need of shelter (REACH, 2018). Therefore, we multiplied the reported total population in need by 0.56 to obtain an estimation for shelter needs. As a result, we obtain two assessment values for shelter needs in each sub-district, which can be utilized to represent demand q_i for shelter at each sub-district of Idleb.

For this case study we assume that a shelter can be opened at every of the 26 nodes in Idleb. Google Maps was used to obtain distances D_{io} between the nodes, i.e., between the centers of the sub-districts. For illustration, we hypothetically set that no more than $\kappa = 10$ shelters can be opened, each with a capacity of $h_o = 100,000$ people, hence, the maximum available shelter capacity is $G = 1,000,000$. Finally, the maximum coverage distance τ is set at 50 kilometers.

5.2 Implementation of the Methodology

In this section, we explain the steps of our methodology implemented to solve the proposed shelter location problem in the Syrian conflict setting.

Step I: Scenario generation As mentioned in Section 4, using easy to define and interpret triangular probability distributions, consisting of a minimum value *min*, maximum value *max* and the most probable value *mode*, can be practical in humanitarian settings to represent uncertainty (Benini et al., 2017). Here, we treat the shelter needs reported in the needs assessment data sets of source #1 (HNO, 2019) and source #2 (REACH, 2018) as the *mode* values, respectively. To capture the uncertainty inherent in the data, we generated the min and max values of the triangular distributions as follows: we chose to set $min = mode * (1 - a)$ and $max = mode * (1 + a)$. Thus, the value $a \in [0, 1]$ corresponds to the confidence associated to the prediction: in the extreme case $a = 0$ there is no uncertainty associated to the prediction, whereas a large value of a implies a low confidence in the prediction, corresponding to high levels of uncertainty. Since the level of uncertainty varies from source to source, but also from sub-district to sub-district, we have chosen to generate the uncertainty parameter a randomly for each source and sub-district prediction. The probability distribution of the uncertainty parameter is given in Table 1.

Table 1: Probability distribution for the uncertainty parameter a

	values for a					
	0.1	0.2	0.3	0.4	0.5	0.6
probability	0.15	0.4	0.15	0.1	0.1	0.1

The sampled uncertainty parameter is represented by the variation of the population in the given sub-district (the *intrinsic uncertainty*), yielding a range in which the true population should lie based on the corresponding source. We have observed several sub-districts where the ranges thus obtained are disjoint, i.e., no value extracted from source # 1 is contained in the range of source #2 and vice versa. This illustrates the potential existence of a second source of *extrinsic uncertainty*: the precision of the estimate may be overestimated.

In order to model such extrinsic uncertainty, we have further widened the ranges by requiring that the ranges are at least touching each other, i.e., if the smallest possible value according to one source is larger than the largest possible value to the other, we increase a until this is no longer the case.

Figure 3 illustrates this approach for the sub-district Bennsh as an example. According to the prediction of source #1, the maximal possible value for this sub-district was about 9,500, whereas the minimal possible value according to source #2 was over 17,000. Therefore, there is a significant source of extrinsic uncertainty, due to the fact that one of the predictions must be off by more than what would be the maximum possible. The original distributions are shown by dashed lines in Figure 3. The distribution obtained after making the adjustment are shown by solid lines.

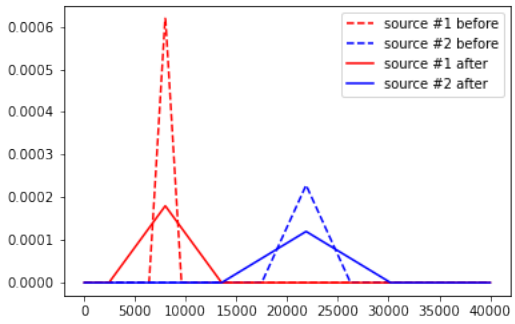


Figure 3: Illustration of the distinction between intrinsic and extrinsic uncertainty, using the Bennsh sub-district. Dashed lines represent distributions pre-adjustment, solid lines the final distributions

Once the value for a has been chosen, 500 scenarios for each source have been generated from the resulting triangular distributions, i.e., we obtain a total of $|\mathcal{S}| = 1,000$ scenarios. Each scenario s can occur with the same probability, i.e., $p^s = 0.001$.

Step II: Opportunity cost distance In the second step of our methodological process, the opportunity cost distances $d(\cdot, \cdot)$ had to be determined. For this purpose, our two-stage stochastic model (7)-(16) was solved for each scenario separately and differences between the corresponding objective values were calculated via (18). In the case where a single scenario is considered, (7)-(16) becomes a deterministic model.

Step III: Cluster generation Using the opportunity cost distance $d(\cdot, \cdot)$ from the previous step, we now have a graph on the set of 1,000 scenarios, where the length of the edge between any two vertices s_1 and s_2 is given by the opportunity cost $d(s_1, s_2)$. This leads us to the graph clustering problem of identifying clusters of vertices such that the edge between any two scenarios from the same scenario is short. Based on the opportunity cost distances in (18), we have grouped the scenario set using the normalised N-Cut algorithm, as mentioned in the third step of our methodology in Section 4.1. In this algorithm, the number of clusters M

is an input parameter that can be chosen. Specifically, we have clustered the graph into 2, 3, ..., 39 clusters and haven chosen the optimal clustering according to the Silhouette score. While this upper bound of 39 may seem to be arbitrary, we have found that as the number of clusters grows above 10, the quality of the clustering decreases rapidly. Therefore, the upper bound does not turn out to be very important.

Step IV: Ambiguity analysis In the last step of the descriptive phase, we analyze the consensus level between sources by determining the decisional level of agreement (19) and the diversity of sources in a cluster (21) based on the previously generated clusters. The corresponding results are shown in Section 5.3.

Step V: Integration through optimization (prescriptive phase) Following the application of the descriptive phase, the integration step involves identifying the *consensus decisions*, which are obtained through optimization (22). The determination of the corresponding weights w_s are explained in the following.

Let us define $z = (z_o : o \in O)$ as a binary vector that includes the shelter opening decisions. We further define $Z = \{z \mid \sum_{o \in O} z_o \leq k, z_o \in \{0,1\}, \forall o \in O\}$ as the set of first-stage constraints. Considering a solution $z \in Z$, we also express the second-stage cost function $\phi(z, q^s, \theta^s) = \max \sum_{o \in O} p_s r_o^s$ for a specific scenario $s \in S$, such that constraints (9)-(13) and (15)-(16) hold. For a given set $\overline{S} \subseteq S$ and the weight values w_s, $s \in \overline{S}$, the integration optimization model in (22) is defined for our case as follows:

$$\max \quad \sum_{s \in \overline{S}} w_s \phi(z, q^s, \theta^s) \tag{26}$$

$$\text{s.t.} \quad z \in Z. \tag{27}$$

Therefore, the consensus decisions, which we denote as $z^\star$ (i.e., the optimal solution for model (26)-(27)), are directly dependent on the choices made regarding the set $\overline{S}$ and how the weights w_s, $s \in \overline{S}$, are fixed.

Regarding our specific application, we present four strategies to fix the set $\overline{S}$ and the weights w_s, $s \in \overline{S}$, in solving model (26)-(27):

1. *Expected value approach*: The expectation is applied over the information based on both sources as the means to integrate. When applied in our case problem, this entails that we define the expected scenario $\overline{s}$ for which the associated parameters are defined as follows: $q^{\overline{s}} = (\overline{q}_i : i \in I)$, where $\overline{q}_i = \sum_{s \in S} p_s q_i^s$, $\forall i \in I$ and $\theta^{\overline{s}} = \sum_{s \in S} p_s \theta^s$. Thus, to obtain the consensus decisions in this case, we fix $\overline{S} = \{\overline{s}\}$ and we set the value $w_{\overline{s}} = 1$. Model (26)-(27) is then solved, and we let $\overline{z}$ define the optimal solution obtained.

2. *Stochastic optimization*: This is the traditional stochastic programming approach, which approximates the stochastic phenomena that is present in the considered problem by generating a set of representative scenarios. In this case, we thus define $\overline{S} = S$ and we set $w_s = \frac{1}{\mid S \mid}$, $\forall s \in \overline{S}$, to account for the fact that the confidence level for all sources is identical (i.e., we thus assume that all scenarios are equiprobable). Model (26)-(27) is then solved and we let $\tilde{z}$ define the optimal solution obtained.

3. *Scenario clustering*: The clusters generated in step III of our methodology are used to perform the ambiguity analysis to assess the level of consistency between the sources regarding the information they are providing. In the present case, we set $\overline{S} = S$ and determine the weights w_s, $s \in \overline{S}$ using equation (25). Model (26)-(27) is then solved and we let $\hat{z}$ define the optimal solution obtained.

4. *Source specific integration*: This approach relies solely on the information provided by the first and second source, respectively. Therefore, we define $\overline{\mathcal{S}} = \mathcal{S}^1$ ($\overline{\mathcal{S}} = \mathcal{S}^2$) and we set $w_s = \frac{1}{\mid \mathcal{S}^1 \mid} \left(w_s = \frac{1}{\mid \mathcal{S}^2 \mid} \right)$,

$\forall s \in \overline{\mathcal{S}}$. The model (26)-(27) is then solved to obtain the optimal solution $z_1^\star$ and $z_2^\star$, respectively. In this case, solution $z_1^\star$ ($z_2^\star$) can be interpreted as the best possible solution if source #1 (#2) is used in the assessment of the needs.

For the case with two data sources, (4) and (5) can be written as:

$$\epsilon_1(z) = \sum_{s \in \mathcal{S}^1} \frac{1}{\mid \mathcal{S}^1 \mid} \phi(z_1^\star, q^s, \theta^s) - \sum_{s \in \mathcal{S}^1} \frac{1}{\mid \mathcal{S}^1 \mid} \phi(z, q^s, \theta^s), \tag{28}$$

$$\epsilon_2(z) = \sum_{s \in \mathcal{S}^2} \frac{1}{\mid S_2 \mid} \phi(z_2^\star, q^s, \theta^s) - \sum_{s \in \mathcal{S}^2} \frac{1}{\mid S_2 \mid} \phi(z, q^s, \theta^s), \tag{29}$$

$$\epsilon(z) = \epsilon_1(z) + \epsilon_2(z). \tag{30}$$

In the following, we use our methodological framework to analyze the ambiguity of both data sources and to evaluate the proposed clustering approach.

5.3 Results and Analysis

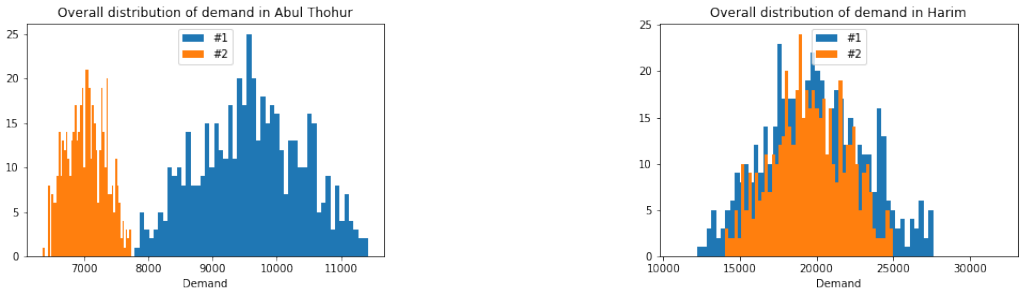
In this section, we apply the steps of our methodology and present results for our case instance that focuses on making shelter location decisions based on multiple needs assessments. Given the scenarios generated in step I of our methodology, step II consists in solving model (7)-(16) for each scenario $s \in S$ to obtain shelter solutions z_o according to (17). We have found that some of the shelters are 'uncontroversial,' in the sense that they are opened either in almost all scenarios or in none of them. Table 2 shows which shelter locations are chosen in more than 90% and fewer than 10% of scenarios overall. For instance, node 2 is chosen for opening a shelter in more than 90% of the scenarios, i.e., independent of the data source. In contrast, shelter locations 10, 12, 13, 16, 17, 18, 20, are almost never part of the solution. For the remaining 18 locations, such generalization for opening or not, cannot be made.

Table 2: Location decisions to open shelters

Decision	Shelter
Almost always open ($> 90\%$)	2
Almost never open ($< 10\%$)	10, 12, 13, 16, 17, 18, 20
'Controversial' shelters	1, 3, 4, 5, 6, 7, 8, 9, 11, 14, 15, 19, 21, 22, 23, 24, 25, 26

The reason for the 'controversial' cases can be found in the distribution of overall demand according to the two sources. In some cases, these predictions are quite far apart. Consider for example the distribution of the overall demand prediction for Abul Thohur, illustrated in Figure 4a. Here, the ranges of estimated values based on the two sources barely overlap. In other words, there is high ambiguity between the two data sources with respect to the prediction of shelter demand, as the sources do not even agree on the range of feasible values.

At the other extreme, there are districts where there is very low ambiguity since the predictions of the two sources almost completely coincide. Consider for example Figure 4b, where the overall demand prediction for Harim is shown.



(a) (b)

Figure 4: Overall demand prediction for Abul Thohur (a) and Harim (b), according to sources #1 and #2

The question arises as to where shelter locations should be opened when demand assessments differ greatly in some cases, e.g. as in Abul Thohur, and most shelter locations are 'controversial' (Table 2). To answer this question, the ambiguity of both data sources has to be analyzed and integrated in the decision-making process.

By implementing step III of the proposed methodology and based on the Silhouette score, the optimal number of clusters is $M = 7$. These clusters are used in the following to perform the ambiguity analysis, as described in step IV of our methodology. As illustrated by Figure 5, the scenarios generated from source #1 split over 5 clusters and the last two clusters consist exclusively of the source #2 scenarios. Two observations can be made. First, source #1 predicts a much higher level of uncertainty than source #2 as it is present in more clusters. Second, the clusters are very homogeneous with respect to the data source from which the scenarios were generated: in all clusters only one data source is present, i.e., there is no diversity of data sources within the clusters and therefore no entropy. This means that in terms of the shelter solution there is a high degree of disagreement between the two data sources.

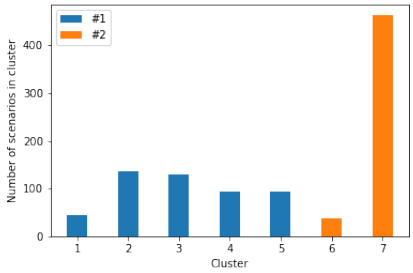


Figure 5: Distribution of scenarios across the clusters

Within each of the seven generated clusters the decision level of agreement (19) is shown in Table 3. A graphical representation of the distribution of opened shelters across the clusters is also given in Figure 6 in the Appendix. According to the results, in clusters consisting of scenarios from source #1, i.e., C_1 - C_5, there is a relatively greater consensus regarding shelter locations than in those from source #2, i.e., C_6 and C_7, resulting in a higher credibility of source 1#. Such analyses allow the decision-maker to understand the

level of ambiguity in the information coming from different sources and its impact on shelter locations. Such insights cannot be gained when traditional stochastic optimization approaches are utilized.

Table 3: Decision level of agreement by cluster C_j

$\Delta(C_1)$	$\Delta(C_2)$	$\Delta(C_3)$	$\Delta(C_4)$	$\Delta(C_5)$	$\Delta(C_6)$	$\Delta(C_7)$
0.3371	0.3505	0.3654	0.3608	0.3209	0.5465	0.5694

Next, we implement step V of our methodology by identifying consensus decisions $z^\star$ for different approaches, i.e., *expected value*, *stochastic optimization*, *scenario clustering* as well as the *source specific integration* described in the last step in Section 5.2. The corresponding shelter solutions are shown in Table 4. Although source #1 and #2 estimate shelter needs for some districts differently, e.g. as in the case of Abul Thohur, shelter solutions $z_1^\star$ and $z_2^\star$ for data source #1 and #2, respectively, have many overlaps. According to Table 4, both solutions coincide for 8 out of 10 possible shelter locations, namely at nodes 2, 3, 21, 22, 23, 24, 25, and 26. In contrast to Table 2, where shelter locations are chosen in a deterministic setting, i.e., for a particular scenario, shelter solutions of source #1 and #2 are similar when uncertainty is taken into account.

Notably, shelter locations chosen by the *expected value* approach have six overlaps with source #1 and #2, whereas the *stochastic* solution has only two overlaps. In contrast to the expected value, the stochastic approach does not try to find the best solution for a specific scenario, but across all scenarios. However, both approaches neglect the ambiguity inherent in the data sources. To account for the underlying ambiguity, the shelter solution for the *clustering* approach has been computed with weights w_s in (25) based on $M = 7$. Due to the lack of data source diversity, i.e., $H(C_j) = 0$ for $j \in M$, the first weight in (23) is $w_j^{(1)} = 0.1733$ for each cluster. According to (24), the second weight is $w_{\#1}^{(2)} = 1.0285$ and $w_{\#2}^{(2)} = 0.387$ for data source #1 and #2, respectively, leading to the final weight $w_s = 0.1782$ for scenarios generated by source #1 and $w_s = 0.0671$ by source #2. Therefore, scenarios coming from the risk-averse source #1 are weighted more than those from source #2, as it is present in more clusters showing its rather "stochastic" attitude. In other words, source #1 predicts a higher level of uncertainty, which can be considered more realistic and is therefore weighted more. Such integrated analysis, i.e., taking into account the impact on the decision problem at hand, reveals which source should be given more weight. As a result, the corresponding shelter solution $\hat{z}$ in Table 4 indicates an overlap more with shelter locations based on source #1 than with #2. Overall, our clustering approach leads to more shelter overlaps with both data sources #1 and #2 than the expected value and the stochastic approach, see $\overline{z}$ and $\tilde{z}$, respectively. The remaining shelter locations, i.e. 6, 17 and 18, were chosen by the clustering approach to hedge against ambiguity and risk. These results show that the proposed methodology can provide an effective means of guiding the decision-maker to reach a consensus decision based on conflicting information from multiple reliable information sources such as experts, and hence addresses an important need in practice as highlighted by humanitarian practitioners (e.g., Benini et al. (2017)).

5.4 Out-of-Sample Tests

Now, we evaluate the objective value obtained by the proposed clustering method compared with respect to the expected value and stochastic approaches. For this purpose, out-of-samples tests were carried out, where

Table 4: Shelter locations for different approaches

Shelter Location	Node	Source #1 $z_1^\star$	Source #2 $z_2^\star$	Expected Value $\bar{z}$	Stochastic $\tilde{z}$	Clustering $\hat{z}$
Abul Thohur	1		x			
Bennsh	2	x	x	x		
Idleb	3	x	x	x		x
Maaret Tamsrin	4			x	x	
Saraqab	5			x	x	
Sarmin	6			x	x	x
Teftnaz	7	x		x		x
Heish	8			x		
Kafr Nobol	9				x	
Khan Shaykun	10					
Ma'arrat An Nu'man	11		x		x	x
Sanjar	12					
Tamanaah	13	x		x		x
Armanaz	14				x	
Dana	15				x	
Harim	16				x	
Kafr Takharim	17				x	x
Qourqeena	18					x
Salqin	19					
Badama	20					
Darkosh	21	x	x		x	x
Janudiyeh	22	x	x			
Jisr-Ash-Shugur	23	x	x			
Ariha	24	x	x			x
Ehsem	25	x	x	x		
Mhambal	26	x	x	x		x

5,000 scenarios were generated for source #1 and #2 each based on the same principles as before and shelter locations from Table 4 were used as an input.

Table 5 shows the gaps (28)-(30) between the objective values of the out-of-sample tests for the expected value, stochastic and clustering approaches and the objective values for source #1 and #2, respectively. For instance, the expected value approach leads to a solution where 1,075 fewer people or families can be accommodated than in the solution based on source #2, i.e. $\epsilon_2(\bar{z}) = 1,075$. Although the shelter solution of the expected value approach has many overlaps with both data sources, see Table 4, it performs worst in terms of the objective value. The number of overlaps alone is no guarantee for a good objective value, as the stochastic shelter solution hardly overlaps with both data sources, but still leads to lower objective gaps. In contrast, our scenario clustering approach provides the lowest gap results, meaning that solution $\hat{z}$ best integrates the information coming from source #1 and #2 while hedging against ambiguity and uncertainty. In particular, it provides the same objective value as source #1 and at the same time can accommodate more people than the other two approaches in the case of source #2. This clustering method can thus support the humanitarian decision-maker to incorporate divergent information of different data sources in a way that higher demand satisfaction can be achieved.

It should be recalled that these numerical experiments only involve the Idleb region and the specific planning of the aid that is provided to service the needs for shelter for the IDP. The proposed clustering method could bring more benefits if applied to multiple affected districts in Syria by considering a broader set of needs for the IDP such as different relief items (e.g., food, hygiene sets, etc.). In this case, it can be expected that further gains will be obtained for both the overall efficiency of the aid that is provided and the hedge that is obtained against the risks stemming from both the ambiguity and the uncertainty in the planning setting.

Table 5: Gaps of objective values for different approaches

Gaps	Source #1 $\epsilon_1(z)$	Source #2 $\epsilon_2(z)$	Total $\epsilon(z)$
Expected Value	84	1075	1159
Stochastic	6	535	541
Clustering	0	402	402

Finally, the out-of-sample tests highlight the overall value of the clustering approach. The assessments of shelter needs provided by source #1 and #2 disagree strongly for some locations. On the one hand, one cannot agree with both sources at the same time. On the other hand, we do not know which of the predictions is closer to the true values. Our clustering approach allows to obtain the smallest gaps while at the same time, integrating the ambiguous information coming from both sources, i.e., the characteristics of the solution provided by our approach are closer to the solutions provided by each source. In this way, a higher level of efficiency is achieved both in terms of the gaps obtained and the solutions that are found. Therefore, a more effective approach is provided that can deal with the ambiguity and the uncertainty that is faced by humanitarian decision-makers.

6 Conclusion

The inherent uncertainty in disaster situations complicates the humanitarian decision-making process. Critical disaster response decisions must be made under significant uncertainty. Furthermore, the complexity of information flow in disaster situations bring significant challenges in making effective decisions. Specifically, different information sources might deliver high-volume data, varying in type and nature, that humanitarian organizations have to gather, analyze and aggregate to estimate the values of important parameters for response such as the needs of the affected people. The available information and estimates from different sources might involve inconsistent elements, which create high levels of ambiguity in decision making. This study takes the first step to present a methodology that can support humanitarian decision-making to analyze the information provided by multiple viable data sources in a systematic and transparent way so that ambiguous information can be transformed into actionable insights and solutions.

We illustrate the proposed approach by focusing on a conflict setting where significant uncertainty and ambiguity may exist in important parameters for making response decisions (such as needs). Specifically, we analyze the estimates of shelter needs in the Syrian civil war derived from two reliable data sources.

Our analyses have revealed a high degree of ambiguity and disagreement between both data sources, as there is a large number of 'controversial' shelter locations and a lack of diversity of data sources within the resulting clusters. Our numerical results show that the proposed methodology better integrates such ambiguous information compared to other common approaches such as the expected value method and stochastic optimization. Specifically, the solutions produced by the new approach are closer to both data sources while achieving greater demand satisfaction, as evidenced by the smaller gaps. Therefore, our newly proposed methodological framework offers humanitarian decision-makers an effective and efficient way to hedge against both ambiguity and uncertainty.

There can be a few future research directions. First, in our case study, our optimization model focuses on a simplified post-disaster shelter location problem for illustration, and the impact of using the proposed methodology in terms of gaps are likely to increase further when more complex models are used. For instance, it would be interesting to evaluate how the clustering method could further improve the decision-making processes when addressing more complex planning problems (e.g., multiple items and periods). Another future research direction would be to explore how to include diverse data types (e.g., unstructured data such as images, audio and video recordings) into the scenario generation step of the proposed methodology. It is crucial to make the best use of all the information provided in today's digitized and data-rich world. Machine learning approaches can be investigated to enhance the proposed methodology to consider diverse data sources, which may have different formats and reliability levels.

References

Abdellaoui M., L'Haridon O., Paraschiv C. Experienced vs. Described Uncertainty: Do We Need Two Prospect Theory Specifications? // Management Science. 2011. 57, 10. 1879–1895.

Altay Nezih, Labonte Melissa. Challenges in humanitarian information management and exchange: evidence from Haiti // Disasters. 2014. 38, s1. 50–72.

Altay Nezih, Pal Raktim. Information diffusion among agents: implications for humanitarian operations // Production and Operations Management. 2014. 23, 6. 1015–1027.

Andres Josh, Wolf Christine T, Cabrero Barros Sergio, Oduor Erick, Nair Rahul, Kjærum Alexander, Thars-gaard Anders Bech, Madsen Bo Schwartz. Scenario-based XAI for Humanitarian Aid Forecasting // Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020. 1–8.

Azizi Shima, Bozkir Cem Deniz Caglar, Trapp Andrew C, Kundakcioglu O Erhun, Kurbanzade Ali Kaan. Aid Allocation for Camp-Based and Urban Refugees with Uncertain Demand and Replenishments // Production and Operations Management. 2021. forthcoming.

Balcik B. Site selection and vehicle routing for post-disaster rapid needs assessment // Transportation Research Part E: Logistics and Transportation Review. 2017. 101. 30–58.

Benini A., Chataigner P., Noumri N., Parham N., Sweeney J., Tax L. The Use of Expert Judgment in Humanitarian Analysis – Theory, Methods, Applications. Geneva: Assessment Capacities Project - ACAPS, 2017.

Birge John R., Louveaux François. Introduction to Stochastic Programming. 2011. Second. (Springer Series in Operations Research and Financial Engineering).

Chamola Vinay, Hassija Vikas, Gupta Sakshi, Goyal Adit, Guizani Mohsen, Sikdar Biplab. Disaster and pandemic management using machine learning: a survey // IEEE Internet of Things Journal. 2020.

Comes Tina, Walle Bartel Van de, Van Wassenhove Luk. The coordination-information bubble in humanitarian response: theoretical foundations and empirical investigations // Production and Operations Management. 2020. 29, 11. 2484–2507.

Crainic Theodor Gabriel, Hewitt Michael, Rei Walter. Scenario grouping in a progressive hedging-based meta-heuristic for stochastic network design // Computers & Operations Research. 2014. 43, 1. 90–99.

Day Jamison M, Melnyk Steven A, Larson Paul D, Davis Edward W, Whybark D Clay. Humanitarian and disaster relief supply chains: a matter of life and death // Journal of Supply Chain Management. 2012. 48, 2. 21–36.

Dong Zhijie Sasha, Meng Lingyu, Christenson Lauren, Fulton Lawrence. Social media information sharing for natural disaster response // Natural Hazards. 2021. 1–28.

Dönmez Zehranaz, Kara Bahar Y, Karsu Özlem, Gama Francisco Saldanha-da. Humanitarian facility location under uncertainty: Critical review and future prospects // Omega. 2021. 102393.

Dyer M., Stougie L. Computational complexity of stochastic programming problems // Mathematical Programming, Series A. 2006. 106, 3. 423–432.

Elçi Özgün, Noyan Nilay. A chance-constrained two-stage stochastic programming model for humanitarian relief network design // Transportation Research Part B: Methodological. 2018. 108. 55–83.

Farahani Reza Zanjirani, Lotfi MM, Baghaian Atefe, Ruiz Rubén, Rezapour Shabnam. Mass casualty management in disaster scene: A systematic review of OR&MS research in humanitarian operations // European Journal of Operational Research. 2020.

Galindo Gina, Batta Rajan. Review of recent developments in OR/MS research in disaster operations management // European Journal of Operational Research. 2013. 230, 2. 201–211.

Grass E., Fischer K. Two-stage stochastic programming in disaster management: A literature survey // Surveys in Operations Research and Management Science. 2016a. 2. 85–100.

Grass Emilia, Fischer Kathrin. Prepositioning of relief items under uncertainty: A classification of modeling and solution approaches for disaster management // Logistics Management. 2016b. 189–202.

Gupta Shivam, Altay Nezih, Luo Zongwei. Big data in humanitarian supply chain management: A review and further research directions // Annals of Operations Research. 2019. 283, 1. 1153–1173.

Gupta Sushil, Starr Martin K, Farahani Reza Zanjirani, Matinrad Niki. Disaster management from a POM perspective: Mapping a new domain // Production and Operations Management. 2016. 25, 10. 1611–1637.

Gutjahr Walter J, Nolz Pamela C. Multicriteria optimization in humanitarian aid // European Journal of Operational Research. 2016. 252, 2. 351–366.

HNO . Humanitarian Needs Overview 2019: Syrian Arab Republic. 2019. Accessed: 23-09-2019.

Hammitt J. K., Shlyakhter A. I. The Expected Value of Information and the Probability of Surprise // Risk Analysis. 2006. 19, 1. 135–152.

Han Jiawei, Pei Jian, Kamber Micheline. Data mining: concepts and techniques. 2011.

Hewitt M., Ortmann J., Rei W. Decision-based scenario clustering for decision-making under uncertainty // Annals of Operations Research. 2021. forthcoming.

Hoffman F. Owen, Hammonds J. S. Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability // Risk Analysis. 1994. 14, 5. 707–712.

Hosseinnezhad D, Saidi-mehrabad M. Data Fusion and Information Transparency in Disaster Chain // International Journal of Innovation, Management and Technology. 2018. 9, 4.

Jahre Marianne, Kembro Joakim, Rezvanian Tina, Ergun Ozlem, Håpnes Svein J, Berling Peter. Integrating supply chains for emergencies and ongoing operations in UNHCR // Journal of Operations Management. 2016. 45. 57–72.

Jain A. K., Dubes R. C. Algorithms for Clustering Data. Upper Saddle River, NJ, USA.: Prentice-Hall, Inc., 1988.

Keutchayan Julien, Ortmann Janosch, Rei Walter. Problem-Driven Scenario Clustering in Stochastic Optimization // arXiv:2106.11717v1. 2021.

Kılcı Fırat, Kara Bahar Yetiş, Bozkaya Burçin. Locating temporary shelter areas after an earthquake: A case for Turkey // European Journal of Operational Research. 2015. 243, 1. 323–332.

Kınay Ömer Burak, Kara Bahar Yetis, Gama Francisco Saldanha-da, Correia Isabel. Modeling the shelter site location problem using chance constraints: A case study for Istanbul // European Journal of Operational Research. 2018. 270, 1. 132–145.

King A. J., Wallace S. W. Modeling with Stochastic Programming. 2012. (Springer Series in Operations Research and Financial Engineering).

Li Hongmin, Caragea Doina, Caragea Cornelia, Herndon Nic. Disaster response aided by tweet classification with a domain adaptation approach // Journal of Contingencies and Crisis Management. 2018. 26, 1. 16–27.

Liberatore F., Pizarro C., Blas C. S., Ortuno M. T., Vitoriano B. Uncertainty in Humanitarian Logistics for Disaster Management: A Review // Decision Aid Models for Disaster Management and Emergencies. 2013. 45–74.

Lloyd S. P. Least squares quantization in PCM. 1957.

Lorca Álvaro, Çelik Melih, Ergun Özlem, Keskinocak Pınar. An optimization-based decision-support tool for post-disaster debris operations // Production and Operations Management. 2017. 26, 6. 1076–1091.

Luxburg Ulrike von. A tutorial on spectral clustering // Statistics and Computing. 2007. 17, 4.

Ni Wenjun, Shu Jia, Song Miao. Location and emergency inventory pre-positioning for disaster response operations: Min-max robust model and a case study of Yushu earthquake // Production and Operations Management. 2018. 27, 1. 160–183.

Noyan N., Balcik B., Atakan S. A stochastic optimization model for designing last mile relief networks // Transportation Science. 2015. 50, 3. 1092–1113.

O'Brien Stephen. This is how we build a stronger, data-driven humanitarian sector. 2017. Accessed: 05-04-2021.

Ofli Ferda, Meier Patrick, Imran Muhammad, Castillo Carlos, Tuia Devis, Rey Nicolas, Briant Julien, Millet Pauline, Reinhard Friedrich, Parkan Matthew, others . Combining human computing and machine learning to make sense of big (aerial) data for disaster response // Big Data. 2016. 4, 1. 47–59.

Paul Jomon A, Zhang Minjiao. Supply location and transportation planning for hurricanes: A two-stage stochastic programming framework // European Journal of Operational Research. 2019. 274, 1. 108–125.

REACH . Situation Overview: Idleb Governorate and Surrounding Areas. 2018. Accessed: 18-07-2021.

Rahimian Hamed, Mehrotra Sanjay. Distributionally robust optimization: A review // arXiv preprint arXiv:1908.05659. 2019.

Raymond Nathaniel, Al Achkar Ziad. Data preparedness: connecting data, decision-making and humanitarian response. 2016. Accessed: 29-10-2021.

Reynard Darcy, Shirgaokar Manish. Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster? // Transportation Research Part D: Transport and Environment. 2019. 77. 449–463.

Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Comput. Appl. Math. 1987. 20. 53–65.

Shi J., Malik J. Normalized Cuts and Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000. 22, 8. 888–905.

Snow Arthur. Ambiguity and the value of information // Journal of Risk and Uncertainty. 2010. 40, 2. 133–145.

Sokat Kezban Yagci, Zhou Rui, Dolinskaya Irina S, Smilowitz Karen, Chan Jennifer. Capturing real-time data in disaster response logistics // Journal of Operations and Supply Chain Management. 2016. 9, 1. 23–54.

Stauffer Jon M, Pedraza-Martinez Alfonso J, Van Wassenhove Luk N. Temporary hubs for the global vehicle supply chain in humanitarian operations // Production and Operations Management. 2016. 25, 2. 192–209.

Swaminathan Jayashankar M. Big data analytics for rapid, impactful, sustained, and efficient (RISE) humanitarian operations // Production and Operations Management. 2018. 27, 9. 1696–1700.

Taylor Kristin, Zarb Stephanie, Jeschke Nathan. Ambiguity, Uncertainty and Implementation // International Review of Public Policy. 2021. 3, 3: 1.

UN Refugee Agency . SYRIA REFUGEE CRISIS. 2021. Accessed: 18-07-2021.

Van Wassenhove Luk N, Besiou Maria. Complex problems with multiple stakeholders: how to bridge the gap between reality and OR/MS? // Journal of Business Economics. 2013. 83, 1. 87–97.

Walle Bartel Van de, Comes Tina. On the nature of information management in complex and natural disasters // Procedia Engineering. 2015. 107. 403–411.

Yáñez-Sandivari Luis, Cortés Cristián E, Rey Pablo A. Humanitarian Logistics and Emergencies Management: New perspectives to a sociotechnical problem and its optimization approach management // International Journal of Disaster Risk Reduction. 2020. 101952.

Yin Shoujun, Jing Runtian. A schematic view of crisis threat assessment // Journal of Contingencies and Crisis Management. 2014. 22, 2. 97–107.

A Illustrative Example

Suppose that there are two data sources: $K = \{1, 2\}$, and that 20 scenarios were generated from each source. The scenario set $\mathcal{S} = \mathcal{S}^1 \cup \mathcal{S}^2$ was grouped into four clusters $C_1, ..., C_4$, with the distribution of scenarios from $\mathcal{S}^1$ and $\mathcal{S}^2$ in each cluster as given in Table 6.

	C_1	C_2	C_3	C_4
source #1	10	2	3	5
source #2	4	4	2	10

Table 6: Distribution of scenarios across clusters and data sources in the illustrative example

In order to calculate the first set of weights, we observe that the proportion of scenarios generated by source #1 are $\frac{5}{7}$, $\frac{1}{3}$, $\frac{3}{5}$ and $\frac{1}{3}$ in clusters 1 through 4 respectively. Thus,

$$H\left(C_1\right) = -\left[\frac{5}{7}\log\left(\frac{5}{7}\right) + \frac{2}{7}\log\left(\frac{2}{7}\right)\right] \approx 0.60,$$

and similarly $H(C_2) = H(C_4) \approx 0.64$ and $H(C_3) \approx 0.67$. Since $|K| = 2$ we have $\lambda_2 = \frac{1}{4}\log(2) \approx 0.173$. This yields

$$w_1^{(1)} \approx 0.773, \quad w_2^{(1)} \approx 0.813, \quad w_3^{(1)} \approx 0.843, \quad w_4^{(1)} \approx 0.813.$$

We now calculate the second set of weights according to (24), where the stochasticity weight for source #1 is

$$\begin{aligned} w_1^{(2)} &= \frac{1}{4} + \frac{1}{|\mathcal{S}^1|^2}\left(\left|\mathcal{S}^1 \cap C_1\right|\left|\mathcal{S}^1 \setminus C_1\right| + \left|\mathcal{S}^1 \cap C_2\right|\left|\mathcal{S}^1 \setminus C_2\right| + \left|\mathcal{S}^1 \cap C_3\right|\left|\mathcal{S}^1 \setminus C_3\right| + \left|\mathcal{S}^1 \cap C_4\right|\left|\mathcal{S}^1 \setminus C_4\right|\right) \\ &= \frac{1}{4} + \frac{10 \cdot 10 + 2 \cdot 18 + 3 \cdot 17 + 5 \cdot 15}{400} = 0.905, \end{aligned}$$

and similarly $w_2^{(2)} = 0.91$.

It remains to calculate the normalization constant W that ensures that the weights are 1 on average:

$$\begin{aligned} W &= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} w_{\iota(s)}^{(1)} w_{\gamma(s)}^{(2)} \\ &= \frac{1}{40}\Big(10w_1^{(1)}w_1^{(2)} + 2w_2^{(1)}w_1^{(2)} + 3w_3^{(1)}w_1^{(2)} + 5w_4^{(1)}w_1^{(2)} \\ &\quad + 4w_1^{(1)}w_2^{(2)} + 4w_2^{(1)}w_2^{(2)} + 2w_3^{(1)}w_2^{(2)} + 10w_4^{(1)}w_2^{(2)}\Big) \approx 0.413. \end{aligned}$$

Putting everything together, we obtain the weights on each scenario according the cluster j it was grouped into and the data source k it was generated from (see Table 7).

	C_1	C_2	C_3	C_4
source #1	1.692	1.376	1.160	0.943
source #2	1.540	0.876	0.420	0.164

Table 7: The final weights in the illustrative example

In conclusion, we observe that the weights defined in (23) and (24) can lead to significant changes in the importance given to the different scenarios. In this example, the weights given to those scenarios coming from source #1 and having been assigned to cluster C_1 are about ten times the weights given to those scenarios coming from source #2 and assigned to cluster C_4. In addition, it can also be seen that all weights associated to the scenarios originating from source #1 are higher, when compared to the weights of the scenarios coming from source #2. This can be explained by the fact that, as mentioned above, clusters with a higher diversity of sources and data sources generating a higher level of stochasticity each receive a greater weight. The implementation of this general approach in an humanitarian environment that involves integrating needs assessment data with shelter location decisions is presented next.

B Proportion of Controversial Shelters

It is interesting to see how the discrepancy between the sources translates into the first-stage decisions z_o to be taken, i.e., which shelters are to be opened. Figure 6 shows the distribution of opened shelters across the clusters for the 'controversial' shelters 1, 3-9, 11, 14, 15, 19, 21-26. In Figure 6, we only show the controversial clusters because the corresponding figures for the others would not be very informative: one would see an empty plot for the 'almost never open' shelters and a plot of bars close to 100% for the 'almost always open' shelter.

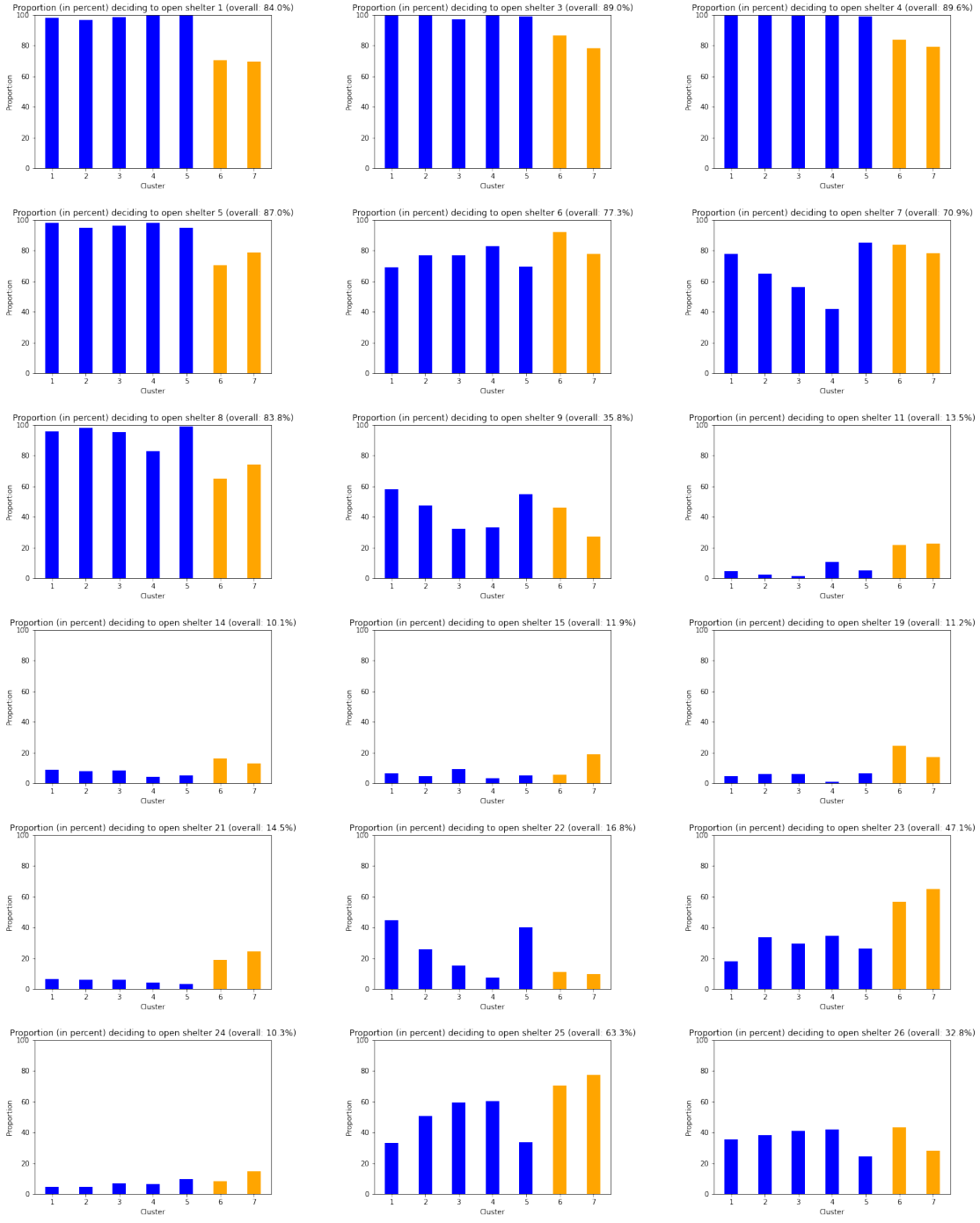


Figure 6: Proportion of scenarios in which the 'controversial' shelters 1, 3, 4, 5, 6, 7, 8, 9, 11, 14, 15, 19, 21, 22, 23, 24, 25, 26 (from left to right, top to bottom) were opened (blue: source #1; yellow: source #2).