Received: 7 December 2021 | Accepted: 9 April 2023

DOI: 10.1111/poms.14018

POMS

ORIGINAL ARTICLE

A machine learning approach to deal with ambiguity in the humanitarian decision-making

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

[1]Business School, University of Mannheim, Mannheim, Germany

[2]GERAD, CRM, and Department of Analytics, Operations and IT, Université du Québec á Montréal, Montréal, Quebec, Canada

[3]Industrial Engineering Department, Ozyegin University, Istanbul, Turkey

[4]CIRRELT, and Department of Analytics, Operations and IT, Université du Québec á Montréal, Montréal, Quebec, Canada

Correspondence
Emilia Grass, Business School, University of Mannheim, 68131 Mannheim, Germany.
Email: grass@uni-mannheim.de

Handling Editor: Sushil Gupta

Abstract

One of the major challenges for humanitarian organizations in response planning is dealing with the inherent ambiguity and uncertainty in disaster situations. The available information that comes from different sources in postdisaster settings may involve missing elements and inconsistencies, which can 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. To the best of our knowledge, the integration of ambiguous information into decision-making by combining a cluster machine learning method with stochastic optimization has not been done before. We illustrate the proposed approach on a realistic case study that focuses on locating shelters to serve internally displaced people (IDP) 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 in making shelter location decisions. 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. With the newly proposed methodology, the decision-maker is able to analyze the degree of ambiguity in the data and the degree of consensus between different data sources to ultimately make better decisions for delivering humanitarian aid.

KEYWORDS
ambiguity, clustering, data aggregation, humanitarian decision-making, needs assessment

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 a natural disaster or during a conflict) can significantly impede effective decision-making processes of humanitarian agencies, which aim to provide timely and sufficient aid (Altay & Labonte, 2014; Day et al., 2012). Specifically, humanitarian agencies have to make decisions under significant uncertainty due to lack of sufficient information on various parameters (e.g., needs, infrastructure conditions) that are critical for disaster response. Moreover, to estimate these parameters, agencies often need to make sense of a large amount of

Accepted by Sushil Gupta, after two revisions.

information with missing and inconsistent elements, which can create high degrees of ambiguity in decision-making. Specifically, ambiguity is defined as "uncertainty about probability, created by missing information that is relevant and could be known" (Snow, 2010). While eliminating ambiguity in postdisaster environments may not be possible, we propose a methodological framework that enhances agencies' capabilities to deal with ambiguity in decision-making.

In postdisaster environments, available information may involve inconsistencies since data can come from a variety of sources (Altay & Labonte, 2014; Day et al., 2012). For instance, postdisaster needs may be estimated by using predisaster information (e.g., governmental statistics) and postdisaster information obtained through various technologies (e.g., aerial images from satellites and drones), and media reports and interviews made by local key informants (such as community leaders). 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 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 postdisaster settings effectively (Comes et al., 2020; O'Brien, 2017; Raymond & Al Achkar, 2016). In this study, we aim to address this important research gap.

Given multiple estimates on a parameter (e.g., 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 the average (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 to use stochastic optimization to support postdisaster decisions such as last-mile relief distribution and shelter location (Dönmez et al., 2021; Liberatore et al., 2013). However, such mathematical aggregation of data without examining its consequences on decision-making can mask the effects and contributions of individual data sources in final decisions (Benini et al., 2017). When the data from different sources are aggregated into a single value or a probability distribution, it is not possible to observe whether the final solution would correspond to a consensus decision if the individual assessments were considered. Thus, one cannot identify which decisions are supported by different estimates, and which ones are significantly affected by the differences among assessments. Decision-makers may also not know which data aggregation techniques to use (computing simple averages or using more sophisticated techniques), and, most importantly, the effects of the chosen aggregation techniques on the final decisions. Therefore, additional information that would reduce such a high level of ambiguity in decision-making would be valuable (Snow, 2010).

Rather than merging data coming from different sources by aggregating before solving a decision-making problem, we 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 postdisaster decision-making, we present an approach based on stochastic optimization and unsupervised machine learning, specifically graph clustering, which aims 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.

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 in 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 initiatives 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 subdistrict 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 subdistrict of Syria. We integrate the 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.

To summarize, the contributions of this paper are as follows:

1. We design a novel and computationally efficient two-phase methodological framework to support the humanitarian decision-making process in a postdisaster setting that includes

(a) a descriptive phase that directly analyzes the levels of information ambiguity stemming from the obtained parameter assessments from the different data sources;
(b) a prescriptive phase that defines a stochastic optimization model by adjusting the relative weights given to each scenario generated in such a way as to reduce the level of ambiguity in the informational context.

To the best of our knowledge, this is the first study to address the problem of ambiguity and divergence of estimates in humanitarian decision-making processes.

2. We demonstrate how the scenario clustering method developed in Hewitt et al. (2022), which relies on the use of a decision-based opportunity loss dissimilarity function to identify patterns in a scenario set, can be generalized and extended to directly analyze the levels of ambiguity that humanitarian decision-makers face when planning operations following a disaster. Specifically, we show that the defined dissimilarity function provides the key to search for clusters of scenarios that exhibit a higher level of decision consensus across multiple data sources. Such clusters then directly reduce the levels of ambiguity in the informational context involved in the planning, which, in turn, provide value to the humanitarian decision-makers. The development of a clustering method in combination with stochastic programming to reduce ambiguity has not yet been done.
3. We illustrate the usefulness and efficiency of our proposed methodological framework using real-world data. Specifically, our case study addresses the integration of needs assessment data with the decisions of locating shelters during the Syrian conflict. We evaluate the quality of our solutions with respect to a naïve approach and a common stochastic optimization model. We also derive insights on the benefits of our proposed approach for the humanitarian decision-makers.

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 decision-making and information management in humanitarian operations (Section 2.1), machine learning and ambiguity (Section 2.2), and shelter location problems (Section 2.3).

2.1 | Humanitarian decision-making and information management

This study is motivated by the need for systematical approaches to facilitate linking information management and decision-making processes, which is a primary challenge in humanitarian environments. There exists a rich literature that present analytical models to address a variety of humanitarian decision-making problems arising in different settings, including transportation planning and fleet management (e.g., Gralla et al., 2016; McCoy & Lee, 2014), inventory and distribution planning (e.g., Azizi et al., 2021; Gallien et al., 2021), prepositioning and network design (e.g., Balcik et al., 2019; Dufour et al., 2018), and postdisaster debris operations (e.g., Lorca et al., 2017). As described by Gralla et al. (2016), humanitarian logisticians must make decisions quickly in emergency situations by using incomplete and large amount of information coming from different sources. Because of the urgency involved, information must be gathered quickly and disseminated to the relevant stakeholders. Several studies highlight the important role of accurate information for humanitarian decision-making and the challenges of information management in disaster contexts (e.g., Altay & Labonte, 2014; Comes et al., 2020; Day et al., 2012; Ergun et al., 2014; Gupta et al., 2016, 2019; P. Shi et al., 2023; Van de Walle & Comes, 2015).

While humanitarian organizations have traditionally suffered from lack of consistent data and information (Starr & Van Wassenhove, 2014), recent advances in technology present major opportunities for leveraging data and information to improve humanitarian operations (Swaminathan, 2018). Humanitarian organizations are increasingly interested in utilizing the benefits of technological innovations in dealing with the complex and dynamic operational environment (Besiou & Van Wassenhove, 2020; Marić et al., 2022; Yoo et al., 2020). However, given that disaster managers are faced with large amounts of information, integration of multiple sources to achieve accurate information for effective decision-making has become an important concern (Gupta et al., 2019). Moreover, no single actor can be the source of all required data in humanitarian environments (Balcik et al., 2010), and different agencies may have different estimates about needs (Ruesch et al., 2022).

Furthermore, in the case of violent conflict situations, it may not be even safe for aid workers to collect data on the ground. Indeed, humanitarian organizations often use external data as it is not feasible for them to collect their own data relevant to plan their response. Therefore, different initiatives (e.g., *Humanitarian Data Exchange* platform, which is managed by United Nations Office for the Coordination of Humanitarian Affairs (OCHA's) Centre for Humanitarian Data) have been launched to ensure access to quality-assured or accurate data and support evidence-based decision-making. How to facilitate such open data by humanitarian agencies operating in the field is also discussed in the literature (e.g., Abuoda et al., 2021; Paulus et al., 2018; Swamy et al., 2019).

Large amounts of information from heterogeneous sources can bring significant challenges for humanitarian organizations that have limited time to make decisions (e.g., Hosseinnezhad & Saidi-mehrabad, 2018). Swaminathan (2018) stresses the need for methods that can effectively synthesize different data streams. Taylor et al. (2021) discuss

that ambiguity and uncertainty are the main reasons for the difference between postdisaster policy formulation and its actual implementation and emphasize the need for new approaches integrating ambiguity, vagueness, and inconsistency. Zagorecki et al. (2013) highlight the importance of applying advanced analysis techniques that creates new knowledge from available data, rather than processing the data in a prescribed manner. 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., Comes et al., 2020; Van Wassenhove & Besiou, 2013). While the existing humanitarian decision-making studies may consider the effects of uncertainties due to data unavailability by using various stochastic and robust optimization approaches, to the best of our knowledge, our study is the first to explicitly analyze the information from different viable sources and integrate them into decision-making.

In this study, we link data processing and decision-making by proposing a 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 (e.g., Ofli et al., 2016; Sokat et al., 2016). Machine learning algorithms have a wide range of applications in disaster management such as identifying damaged buildings, detecting victim locations, predicting the behavior of crowds, assessing risks, and making predictions about disaster occurrences such as floods or fires (e.g., see reviews by Chamola et al., 2020; Linardos et al., 2022; Sun et al., 2020; Zagorecki et al., 2013). The existing techniques mostly focus on developing methodologies to utilize various sources of data to make predictions for informed decision-making in disaster management. However, to the best of our knowledge, there exists no study that utilizes machine learning techniques to integrate estimates from different data sources into decision-making by analyzing and reducing ambiguity, which we address in this study by presenting a novel method.

2.2 | Machine learning and ambiguity

There is a vast literature on combining machine learning and optimization in general: for example, Bengio et al. (2021) and Vesselinova et al. (2020) for supervised approaches and Mazyavkina et al. (2021) for reinforcement learning applied to optimization. Applications of machine learning also appear with increasing frequency in humanitarian logistics, for example, Chamola et al. (2020). In the machine learning literature, the term "ambiguity" appears sometimes (e.g., Ghysels et al., 2021), but it has a different meaning in that context. There, ambiguity is defined in terms of inaccuracy in the data. This presupposes the existence of a ground truth from which the dataset diverges, which does not correspond to our meaning of the term: we do not assume that the ground truth is knowable. To the best of our knowledge, our specific contribution of reducing ambiguity through unsupervised machine learning has not previously been tackled.

Several review papers (e.g., Grass & Fischer, 2016; Gutjahr & Nolz, 2016) show that discrete scenarios are most often used to capture the uncertainties in disaster contexts. There are two general 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. (See Appendix A for an overview on scenario clustering approaches in the Supporting Information.) 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). Our approach uses and extends the methodology of Hewitt et al. (2022) by analyzing the level of decision agreement among scenarios and integrating these scenarios through optimization to reach a consensus decision. Note that the approach of Hewitt et al. (2022) alone is not set up to analyze ambiguity.

2.3 | Shelter location problems

In this study, we propose an integrated data aggregation and decision-making methodology, which is illustrated in a postdisaster setting that focuses on linking the needs assessment data and shelter location decisions during a complex emergency. Both postdisaster needs assessment planning and shelter location problems are widely studied in different humanitarian contexts (e.g., see the reviews by Farahani et al., 2020; Galindo & Batta, 2013). While the assessment information may highly affect the design and management of relief operations, existing studies usually consider data analysis and decision-making 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., Lorca et al., 2017; Stauffer et al., 2016). 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; Kınay et al., 2018; Ni et al., 2018). Given that location decisions are extremely impeded by the high degree of uncertainty inherent in disaster and crisis situations, stochastic optimization techniques are widely utilized (Dönmez et al., 2021). Specifically, two-stage stochastic models have been often used to model uncertainty, which consists 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 exists a high level of uncertainty regarding needs in the affected region. We consider a two-stage stochastic model to locate shelters with limited capacities by exploring how ambiguous needs assessment information can be integrated into the decision-making. Note that robust optimization, particularly 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 and 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.

As discussed in Dönmez et al. (2021), shelter location decisions, which are widely addressed in the literature, are made under significant demand uncertainty (e.g., Kınay et al., 2018; Ozbay et al., 2019). The demand scenarios in the existing papers are often generated based on available data sources (i.e., historical data) by using various methods that rely on different assumptions. That is, there exist no standard datasets and methods followed to generate scenarios based on data. In this paper, we address an important concern that has been raised by practitioners (e.g., Benini et al., 2017), but not been addressed by the studies that use scenario-based approaches in shelter location problems or other humanitarian logistics problems, which is dealing with the ambiguity that may be caused by multiple reliable data sources related to uncertain parameters for disaster response. We illustrate the benefits of the proposed approach by a case study developed with real data from the Syrian conflict.

In summary, this study contributes to the literature by developing a new methodology that links information processing with decision-making in a postdisaster environment that involves uncertainty and ambiguity and presenting the benefits of the proposed approach in 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. Although the proposed approach is illustrated with 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:

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

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 and 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. In the disaster context, scenarios can include the information on, for example, the specific disaster type, its extent, demand for aid, etc. The following discrete probability measure is obtained:

$$P_S = \sum_{s \in S} p_s \delta_s, \tag{2}$$

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:

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

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 [*Step* 1: generate a set of scenarios; *Step* 2: solve the resulting approximated problem (3)], 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:

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

The 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:

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

for some weights ($\lambda_k \geq 0 : k \in K$). These weights represent the relative importance of each source K in the gap. For example, when all sources are considered to be equally reliable (or when there is no information about the reliability), one can choose $\lambda_k = 1$ for all $k \in K$, thus giving each source the same weight in the gap calculation.

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 postdisaster 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 to 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, for example, a civil war as in our case, where it is difficult to obtain

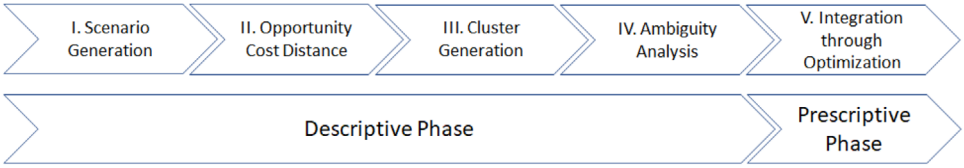

FIGURE 1 General methodological process. [Color figure can be viewed at wileyonlinelibrary.com]

accurate information. For our case study in Section 5.2, we use two data sources, that is, $|K| = 2$, that collect information to assess shelter needs in crisis-affected regions based on different surveys made in the same district in close periods. Depending on the disaster scale, each scenario refers to the number of IDPs in need of shelter. Our network consists of nodes, for example, 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, that is, 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 Appendix B in the Supporting Information is an adopted and simplified version of the one proposed by Noyan et al. (2015).

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 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. In the latter case, the range of values can be used to define probabilistic measures; for example, via 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 general 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 completion, knowledge is obtained on the unknown contextual information of the problem and on the level of overall decision agreement between the models generated from the data sources.

The second phase of our framework is dedicated to the use of this 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, which involve five steps. The descriptive phase is explained in Section 4.1, while the prescriptive phase is presented in Section 4.2. Step I generates sets of scenarios that represent the assessments provided by each data source, following the general stochastic programming approach. Step II calculates the opportunity cost between scenarios in order to quantify the error of predicting the wrong one. Step III identifies groups of scenarios that are close to each other with respect to the opportunity cost defined in Step II. Steps II and III apply a version of the clustering methodology of Hewitt et al. (2022) adapted to our setting. Finally, Steps IV (ambiguity analysis) and V (integration through optimization) are new and make up the key methodological innovations of this paper. Overall, these last two steps lead to defining an optimization model which, once solved, provides us with the consensus solution.

4.1 | Descriptive phase

Following Figure 1, the descriptive phase consists of performing the following four steps: scenario generation, opportunity cost distance computation, 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 dataset, 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 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 5.1.1 and Appendix E for more details in the Supporting Information). Let us recall that we denote by $\mathbb{P}^k$ the source-specific probability distribution associated with 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, that is, 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 and 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

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

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, so

$$|\mathcal{S}| = \sum_{k \in K} |\mathcal{S}^k| = \sum_{k \in K} N_k. \tag{7}$$

Each scenario $s \in \mathcal{S}$ is assigned a probability $p_s \geq 0$ of occurring. Implicit in these probabilities is a weighting of different sources: since scenario sets generated by each data source are disjoint, we can compute the probability of a scenario generated by a given source $k \in K$ to be observed:

$$P_k = \sum_{s \in \mathcal{S}^k} p_s \qquad \text{for } k \in K. \tag{8}$$

The relative values of P_k give a weight of the data source k, representing our confidence in each source. When we are equally confident in all data sources, or when no information is available about their reliability, we can choose the probabilities p_s such that $P_k = \frac{1}{|K|}$. That way, equal weight is associated with each source. One way of achieving this is to generate the same number of scenarios from each source and then to assign equal probabilities to all scenarios: $N_1 = \cdots = N_{|K|}$ and $p_s = \frac{1}{|\mathcal{S}|}$ for all $s \in \mathcal{S}$.

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 $\xi_s, \forall s \in \mathcal{S}$, 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:

$$x\left(s_i^k\right) = \arg\max_{x \in A} \phi\left(x, \xi_{s_i^k}\right), \quad i = 1, \dots, N_k. \tag{9}$$

These solutions can be understood as follows: if one were somehow certain that scenario s_i^k will occur then the solution $x(s_i^k)$, obtained by solving the problem (9) using the predicted scenario s_i^k, will be implemented. Each data source $k \in K$ is associated with the following solution set:

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

and the overall set of all such solutions is thus denoted as

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

We now apply a notion of distance between scenarios, called *opportunity cost distance* that was first introduced in Hewitt et al. (2022). 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 with 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

$$d(s_1, s_2) = \phi(x(s_2), \xi_{s_2}) - \phi(x(s_1), \xi_{s_2}) + \phi(x(s_1), \xi_{s_1}) - \phi(x(s_2), \xi_{s_1}). \tag{10}$$

An opportunity cost distance matrix is then obtained by calculating the distance values using Equation (10) for all

scenario pairs in the overall set (i.e., compute $d(s_1, s_2)$, $\forall s_1, s_2 \in \mathcal{S}$).

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, for example, J. Shi and Malik (2000) and von Luxburg (2007). In the present case, we choose the normalized N-Cut algorithm (Hewitt et al., 2022; J. Shi & Malik, 2000), 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 (10), 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.

We will choose M so as to maximize a particular notion of clustering quality called the *Silhouette score*, which measures how close each scenario is to other members of its own cluster, compared to its distance to other clusters (Rousseeuw, 1987). In the literature, the *elbow method* (Bishop & Nasrabadi, 2006) is sometimes used. In this work, we prefer the Silhouette score because it takes into account both intercluster and intracluster distances. Moreover, the elbow method requires a subjective choice made by the modeler and is therefore less reproducible (Ketchen & Shook, 1996).

Step IV: Ambiguity analysis

This step is dedicated to analyzing the obtained clusters with a focus on diagnosing the level of decision agreement among the scenarios and data sources. We begin by identifying the level of agreement between data sources in terms of the 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

$$\Delta(U) = \frac{4}{|U|^2} \sum_{s_1, s_2 \in U} \Delta(x(s_1), x(s_2)), \tag{11}$$

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, that is, $\Delta(C_j)$ for $j = 1, \ldots, M$. By computing $\Delta(S^k)$, we can also measure the variance of the information obtained from one data source $k \in K$, that is, to what extent the different scenarios generated from k lead to the same solutions (or decisions). Note that at this step, the Hamming distance is used to analyze the agreement or disagreement between different solutions. In the next and last step of our methodology, we incorporate these insights about the agreement level to determine a consensus optimization problem whose solution is evaluated by the epsilon function in (5). (See Appendix C.1 for an example in the Supporting Information.)

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. That is, 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 (find solutions to the problem) that are similar (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:

$$\pi_k(C_j) = \frac{|C_j \cap S^k|}{|C_j|}. \tag{12}$$

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

$$H(C_j) = -\sum_{k \in K} \pi_k(C_j) \log(\pi_k(C_j)), \tag{13}$$

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. (See Appendix C.2 for an example in the Supporting Information.)

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 $\overline{S}$ of the scenario set and weights w_s for each scenario $s \in \overline{S}$, based on the metrics defined above. As a way of formalizing the problem expressed in (6), we define a *consensus solution* as follows:

$$x^\star = \arg\max_{x \in A} \sum_{s \in \overline{S}} w_s \phi(x, \xi_s). \tag{14}$$

This raises questions when formulating problem (14): Which scenarios should be included in $\overline{S}$, and how should the weights w_s be defined?

Regarding the choice of $\overline{S}$, we could include all scenarios: $\overline{S} = S$. 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 S might be very large, and considering the complexity involved in computing the value function ϕ, solving problem (14) 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 (14). As proposed in Hewitt et al. (2022), 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, that is, choosing $\overline{S} \subset S$, naturally leads to an approximation error with respect to using the full set S when searching for a consensus solution (14). That being said, as numerically illustrated in Hewitt et al. (2022), 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 $s \in S$ 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 $\overline{S}$, 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, that is, the weight only depends on the cluster index $j \in \{1, \dots, M\}$ and defined as follows (recalling the definition of H from (13)):

$$w_j^{(1)} = \lambda_K + H(C_j), \quad \text{where} \quad \lambda_K = \frac{\log(|K|)}{4}. \tag{15}$$

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 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, that is, the weight is higher if the source k is more evenly represented across the clusters. More formally, let $\iota: S \longrightarrow \{1, \dots, M\}$ denote the function that maps each scenario s to the index $\iota(s)$ of the cluster to which it belongs, that is, so that $s \in C_{\iota(s)}$. Then

$$w_k^{(2)} = \frac{1}{4} + \frac{1}{|S^k|^2} \sum_{j=1}^{M} |C_j \cap S^k| |S^k \setminus C_j|. \tag{16}$$

The term $\frac{1}{4}$ avoids zero weights if 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 with the scenario probability. 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 $s \in S$ as

$$w_s = \frac{w_{\iota(s)}^{(1)} w_{\gamma(s)}^{(2)}}{W} p_s, \quad \text{where} \quad W = \frac{1}{|S|} \sum_{s \in S} w_{\iota(s)}^{(1)} w_{\gamma(s)}^{(2)} p_s \tag{17}$$

and $\gamma(s) \in K$ denotes the data source from which scenario s was generated, that is, $s \in S^{\gamma(s)}$. The normalization constant

W ensures that the average of the weights is equal to 1:

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

Illustrative examples can be found in Appendixes C.3 and C.4 in the Supporting Information.

5 | NUMERICAL STUDIES

In this section, we present numerical studies 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 IDPs in Idleb. We generate synthetic assessment data and analyze our approach in Section 5.1. In Section 5.2, we focus on real assessments provided by two humanitarian initiatives active on the ground.

5.1 | Synthetic tests

Syria has been at civil war since 2011, which has led to millions of casualties and displaced people (UN Refugee Agency, 2021). For the studies, we focus on the assessments of Idleb district, which is located in the 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. To illustrate our approach, we focus on people in need of shelter in Idleb and use the shelter location model given in Appendix B in the Supporting Information. In our synthetic tests, we suppose that there are five fictitious data sources, denoted by #1 to #5, from which we randomly generate shelter needs assessments. We further distinguish between the "close," "medium," and "wide" cases and simulate from each. The "close" case means that the shelter demand estimations provided by the five sources are rather similar, while in the "wide" case they are far apart. All remaining parameters are presented in Section 5.2.1.

5.1.1 | Implementation of the methodology

In this section, we explain each step of our methodology.

Step I: Scenario generation
We randomly generate 1000 scenarios, that is, 200 from each fictitious source. Each scenario s can occur with the same probability, that is, $p^s = 0.001$, and represents estimates of shelter demand.

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 (22)–(31) in Appendix B in the Supporting Information was solved for each scenario separately and differences between the corresponding objective values were calculated via (10). In the case where a single scenario is considered, (22)–(31) 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 1000 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 (10), we have grouped the scenario set using the normalized 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 have 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 (11) and the diversity of sources in a cluster (13) based on the previously generated clusters.

Step V: Integration through optimization (prescriptive phase)
The integration step involves identifying the *consensus decisions*, which are obtained through optimization (14). 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 $Z = \{z \mid \sum_{o\in O} z_o \le \kappa, z_o \in \{0,1\}, \forall o \in O\}$ 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 (24)–(28) and (30) and (31) of the optimization model in Appendix B (in the Supporting Information) hold. For a given set $\overline{S} \subseteq S$ and the weight values w_s, $s \in \overline{S}$, the integration optimization model in (14) is defined for our case as follows:

$$\max \sum_{s\in\overline{S}} w_s \phi(z, q^s, \theta^s) \tag{18}$$

$$\text{s.t. } z \in Z. \tag{19}$$

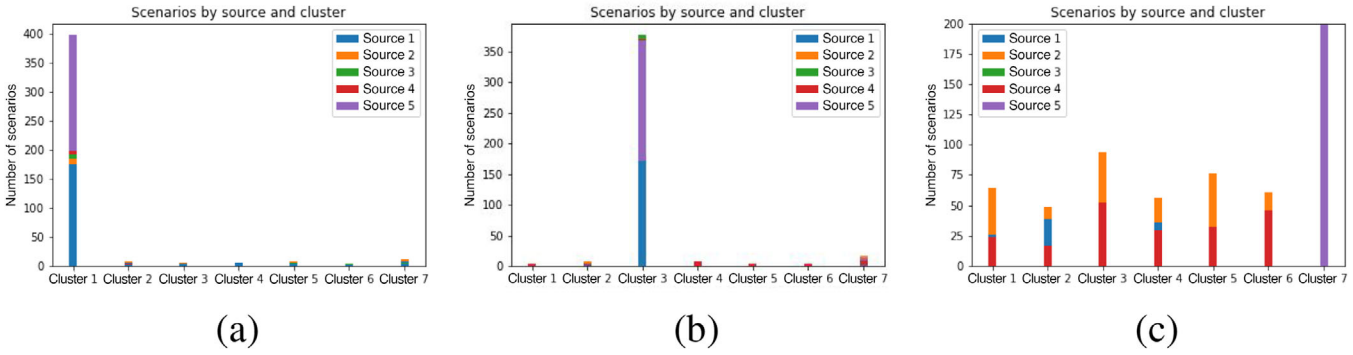

FIGURE 2 Distribution of scenarios across the clusters for case (a) "close," (b) "medium," and (c) "wide." [Color figure can be viewed at wileyonlinelibrary.com]

The objective function (18) maximizes the weighted expected number of accommodated people over all scenarios from $\overline{S}$, so that a maximum of κ shelters can be opened in the first stage (19). Therefore, the consensus decisions, which we denote as $z^\star$ (i.e., the optimal solution for model (18)–(19)), 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 (18)–(19):

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 (18)–(19) 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 $w_s = \frac{1}{|S|}$, $\forall s \in \overline{S}$, 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 (18)–(19) 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 (17). Model (18)–(19) 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 five sources. Therefore, we define $\overline{S} = S^k$ and we set $w_s = \frac{1}{|S^k|}$, $\forall s \in \overline{S}$ and $k = 1, \ldots 5$. Model (18)–(19) is then solved to obtain the optimal solution $z_k^\star$. In this case, solution $z_k^\star$ can be interpreted as the best possible solution if source k is used in the assessment of shelter needs.

As we have no information about the relative reliability of the five sources, we weight them equally. This corresponds to choosing $\lambda_k = 1$ for each data source k in (5). In particular, (4) and (5) can then be written as

$$\epsilon_k(z) = \sum_{s \in S^k} \frac{1}{|S^k|} \phi(z_k^\star, q^s, \theta^s) - \sum_{s \in S^k} \frac{1}{|S^k|} \phi(z, q^s, \theta^s), \tag{20}$$

$$\epsilon(z) = \sum_{k \in K} \epsilon_k(z). \tag{21}$$

5.1.2 | Results and analysis

In this section, we apply the steps of our methodology and present results for our cases that focus on making shelter location decisions based on multiple needs assessments. Based on the scenarios generated in Step I and shelter solutions z_o obtained in Step II, the optimal number of clusters in Step III is $M = 7$ (according to the Silhouette score) for all three cases, with the respective distributions shown in Figure 2. In contrast to the "close" and "medium" cases, where most scenarios are present in one cluster, in the "wide" case they are rather evenly distributed across the first six clusters. This observation is confirmed by the entropy, where the "medium" case has the highest spread of entropy values and therefore a particularly high level of ambiguity. (See Appendix D for a discussion on the ambiguity analysis (Step IV) and the resulting weights (Step V) in the Supporting Information.)

We now implement Step V of our methodology by identifying consensus decisions $z^\star$ for the approaches: *expected value*, *scenario clustering,* and the *source-specific integration* described in Section 5.1.1. In order to evaluate the performance of our clustering approach, Table 1 summarizes the gaps according to (20) and (21) for the "close," "medium," and "wide" cases. For all three cases, our proposed method

TABLE 1 Comparison of gaps between the expected value and the new clustering approach for the "close," "medium," and "wide" cases.

Case	Approach	Source					Total gap
		#1	#2	#3	#4	#5	
Close	Expected value	3097	4952	5657	4790	6356	24,853
	Clustering	130	119	123	190	172	734
Medium	Expected value	9420	11,438	10,036	10,296	11,908	53,098
	Clustering	93	33	130	76	50	381
Wide	Expected value	775	1740	1666	920	901	6003
	Clustering	93	30	199	190	86	598

outperforms the *expected value approach*, resulting in significantly lower gaps. Starting with the "close" case, the total gap of the expected value method is about 34 times higher. If the estimates provided by the sources are further apart, that is, "medium" case scenarios, then this total gap is even larger. One would expect these gaps to increase even further for the "wide" case. However, the situation here is somewhat different. Although estimates from sources diverge widely, the ambiguity analysis has revealed a similar agreement level but a higher entropy on average. In other words, while sources vary in their estimates of shelter needs, there is a fair consensus on shelter locations. Therefore, shelter solutions based on the clustering approach are less dissimilar to the expected value approach when compared to the "close" and "medium" cases. That being said, one still observes that the use of the clustering approach leads to a solution that is noticeably closer to each source-specific solution found for the "wide" case setting.

These results confirm our intuition that our newly proposed methodological framework is particularly beneficial when there is a wider range of source estimates, less consensus on the solutions, and less diversity of sources within the clusters. Given that humanitarian environments highly exhibit these characteristics due to the large number and diversity of primary and secondary data sources, as highlighted by the literature, the proposed methodology can be highly useful for bridging humanitarian information management and decision-making.

5.2 | Idleb case study

We give background information on the needs assessment data provided by two different sources (Section 5.2.1) and conclude with the corresponding results (Sections 5.2.2 and 5.2.3).

5.2.1 | Case dataset

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 (cleaned, combined, cross-checked with secondary sources) and the sector-specific needs (shelter, nutrition, etc.) in each district are published publicly.

We focus on two major assessment datasets, which are made publicly available by two humanitarian initiatives, the Humanitarian Needs Overview (HNO) and REACH. There are other initiatives and organizations in Syria that focus on collecting and disseminating assessment data at different capacities (e.g., The International Organization for Migration, United Nations High Commissioner for Refugees). Because HNO and REACH data are available for the same period of time for shelter needs in the our focus district Idleb, we consider two sources.

HNO, a joint effort by the United Nations (UN) and its partners to assess the humanitarian situation in Syria, provides a consolidated and comprehensive dataset including estimates on the number of people in need for different types of relief in each district of Syria. We consider the nationwide needs assessment of HNO conducted for 6322 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.

REACH, a nonprofit organization that aims to support humanitarian response through better information management, provides needs assessment data for the estimated total number of people residing in a district and the percentage of people requiring different types of supplies, for example, water, food. and shelters. We consider the assessment dataset 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 which provides estimates on the humanitarian needs.

While source #1 provides the estimated number of people requiring shelter in detail, source #2 provides an aggregate estimate according to which about 56% of local people are in need of shelter (REACH, 2018). We, therefore, multiplied the reported total population by 0.56 to obtain an estimation for shelter needs. We obtain two assessment values for

TABLE 2 Location decisions to open shelters.

Decision	Shelter
Almost always open (>80%)	3, 4
Almost never open (<20%)	9, 10, 12, 14, 16, 17, 19, 20, 25, 26
"Controversial" shelters	1, 2, 5, 6, 7, 8, 11, 13, 15, 18, 21, 22, 23, 24

shelter needs in each subdistrict, which can be used to represent demand for shelter at each subdistrict of Idleb.

As mentioned in Section 4, 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 datasets of source #1 (HNO, 2019) and source #2 (REACH, 2018) as the *mode* values, respectively. Based on generated triangular distributions discussed in Appendix E in the Supporting Information, 500 scenarios for each source have been designed, that is, we obtain a total of $|S| = 1000$ scenarios. Since HNO and REACH are both supported by the UN and are the most widely used needs assessment sources, we assign the same level of confidence to both sources, that is, equal weights P_k in (8) and λ_k in (5) for both sources.

Idleb consists of 26 subdistricts, and we assume that a shelter can be opened at every of these subdistricts. Google Maps was used to obtain distances between the centers of the subdistricts. Since 2016, the OCHA organization has provided monthly information on the locations of shelters in Idleb and Aleppo, the types of shelters, and the number of IDPs accommodated. According to OCHA (2018), shelters were opened in 10 subdistricts of Idleb in August 2018, accommodating an average of 80,000 IPDs. Based on this information, we set that no more than 10 shelters can be opened, each with an average capacity of 80,000 people; hence, the maximum available shelter capacity is 800,000. As indicated by the Syria Needs Analysis Report (ACAPS, 2014), most IDPs have fled to their neighboring districts. Therefore, the maximum coverage distance is set to 35 km, which is the average distance between two neighboring districts.

5.2.2 | Results and analysis

In the following, we again follow the steps presented in Section 5.1.1. Given the scenarios generated in Step I of our methodology, Step II consists in solving model (22)–(31) (Appendix B in the Supporting Information) for each scenario $s \in S$ to obtain shelter solutions z_o according to (9). 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 80% and fewer than 20% of scenarios overall. For instance, nodes 3 and 4 are chosen for opening a shelter in more than 80% of the scenarios, that is, independent of the data source. In contrast, shelter locations 9, 10, 12, 14, 16, 17, 19, 20, 25, and 26 are almost never part of the solution. For the remaining 15 locations, such generalization for opening or not, cannot be made.

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 3a. 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 3b, where the overall demand prediction for Harim is shown. The question arises as to where shelter locations should be opened when demand assessments differ greatly in some cases, for example, 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 $M = 9$. These clusters are used in the following to perform the ambiguity analysis, as described in Step IV of our methodology. As illustrated by Figure 4, the scenarios generated from source #1 split over nine clusters and the last cluster consists of the source #2 scenarios and only one #1 scenario. 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 (except cluster 9 having one #1 scenario), that is, there is no diversity of data sources within the clusters and therefore no entropy for clusters 1–8 and a negligible entropy of 0.0144 for cluster 9. This means that in terms of the shelter solution there is a high degree of disagreement between the two data sources.

Within each of the nine generated clusters, the decision level of agreement (11) is shown in Table 3. A graphical representation of the distribution of opened shelters across the clusters is also given in Figure 7 in Appendix F in the Supporting Information. According to the results, in clusters consisting of scenarios from source #1, that is, C_1–C_8, there is relatively less consensus regarding shelter locations than in those from source #2, that is, C_9, resulting in a higher credibility of source #2. 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.

The shelter solutions for the different approaches described in Step V in Section 5.1.1 are shown in Table 4. Column "Actual" indicates where shelters were actually opened in August 2018 in Idleb (OCHA, 2018). As sources #1 and #2 estimate shelter needs for some districts differently, for

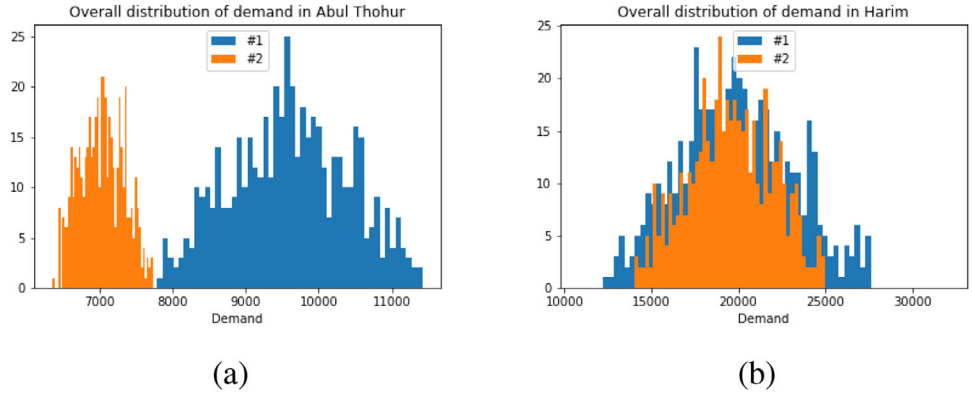

FIGURE 3 Overall demand prediction for Abul Thohur (a) and Harim (b), according to sources #1 and #2. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Decision level of agreement by cluster C_j.

$\Delta(C_1)$	$\Delta(C_2)$	$\Delta(C_3)$	$\Delta(C_4)$	$\Delta(C_5)$	$\Delta(C_6)$	$\Delta(C_7)$	$\Delta(C_8)$	$\Delta(C_9)$
0.6981	0.7033	0.6811	0.6627	0.7109	0.7521	0.6958	0.7256	0.3868

TABLE 4 Shelter locations for different approaches.

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

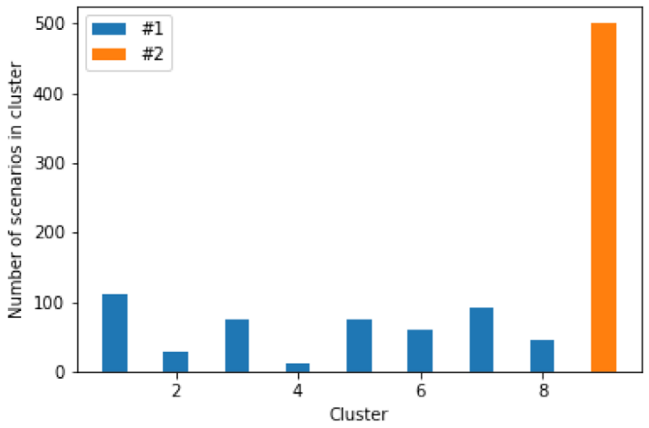

FIGURE 4 Distribution of scenarios across the clusters. [Color figure can be viewed at wileyonlinelibrary.com]

example, as in the case of Abul Thohur, shelter solutions $z_1^\star$ and $z_2^\star$ for data source #1 and #2, respectively, have only three out of 10 overlaps, namely at nodes 11 (Ma'aarat An Nu'man), 15 (Dana), and 21 (Darkosh). These are also the subdistricts where shelters have been opened.

Notably, shelter locations chosen by the *expected value* and the *stochastic* approach have more overlaps with sources #1 and #2 than our *clustering* approach. To account for the underlying ambiguity, the shelter solution for the *clustering* approach has been computed with weights w_s in (17) based on $M = 9$. Due to the lack of data source diversity, that is, $H(C_j) = 0$ for $j = 1, \ldots, 8$, the first weight in (15) is $w_j^{(1)} = 0.1733$ for clusters 1–8 and $w_9^{(1)} = 0.1877$ for cluster 9. According to (16), the second weight is $w_{\#1}^{(2)} = 1.0954$ and $w_{\#2}^{(2)} = 0.25$ for data sources #1 and #2, respectively, leading to the final weight $w_s = 1.6034$ for scenarios generated by source #1 and $w_s = 0.3963$ 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, that is, 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 two overlaps more with shelter locations based on source #1 than with #2. Overall, our clustering approach leads to seven shelter overlaps with data sources #1 and #2, whereas the remaining shelter locations, that is, 12, 16, and 19, 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).

TABLE 5 Gaps of objective values for different approaches.

	Source #1	Source #2	Total
Gaps	$\epsilon_1(z)$	$\epsilon_2(z)$	$\epsilon(z)$
Expected value	1	22,789	22,790
Stochastic	6	583	589
Clustering	0	68	68

5.2.3 | 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 5000 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 (20) and (21) between the objective values of the out-of-sample tests for the different approaches. For instance, solution of the expected value accommodates 22,789 fewer people than the solution based on source #2. Though the shelter solution of the expected value approach has many overlaps with both data sources, it performs worst in terms of the objective value. The number of overlaps alone is no guarantee for a good objective value.

According to Table 5, the stochastic approach results in a smaller gap for source #2: $\epsilon_1(\tilde{z}) = 583$ versus $\epsilon_1(\bar{z}) = 22,789$, but at the expense of a higher gap for source #1, that is, $\epsilon_2(\tilde{z}) = 6$ versus $\epsilon_2(\bar{z}) = 1$. In contrast, our scenario clustering approach provides the lowest gap results, meaning that solution $\hat{z}$ best integrates the information coming from sources #1 and #2 while hedging against ambiguity and uncertainty. It provides the same objective value as source #1 and can accommodate more people than the other two approaches in the case of source #2. In summary, our method can support the humanitarian decision-maker to incorporate divergent information of different data sources in a way that higher demand satisfaction can be achieved.

Recall 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 IDPs. 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.

The out-of-sample tests highlight the overall value of the clustering approach: the shelter needs assessments provided by sources #1 and #2 disagree strongly for some locations. One cannot agree with both sources at the same time, but we do not know which of the predictions is closer to the true values. Our clustering approach obtains the smallest gaps while

at the same time integrating the ambiguous information. That is, the characteristics of our solution are closer to the solutions provided by each source. In this way, a higher level of efficiency is achieved terms of the gaps obtained and the solutions. We have provided a more effective approach 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 brings 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. We present the first methodological framework 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 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. This is also confirmed by additional synthetic tests for different levels of estimation ranges, ambiguity, and consensus, where the clustering method outperforms the naïve approach in every single case. These tests have revealed that our clustering method is particularly advantageous when there is a wider range of source estimates, less consensus on the solutions, and less diversity of sources within the clusters. Overall, our results suggest that our clustering approach is likely to be particularly valuable in cases with a high degree of ambiguity and can therefore offer humanitarian decision-makers an effective and efficient way to hedge against both ambiguity and uncertainty.

Our work suggests several future research directions. Our optimization model focuses on a simplified shelter location problem for illustration. The impact of using the proposed methodology in terms of gaps is likely to increase further when more complex models are used. It would be interesting to evaluate this improvement when addressing more complex planning problems (e.g., multiple items and periods). In this paper, we have focused solely on a classical stochastic optimization approach, minimizing expected cost. The methodology of this paper can also be translated to variants of stochastic programming with alternative objective functions and also robust optimization framework (e.g., using our analytical method to reduce the size of the ambiguity sets in distributionally robust models), which can be addressed by future studies.

ACKNOWLEDGMENTS

Open access funding enabled and organized by Projekt DEAL.

ORCID

Emilia Grass https://orcid.org/0000-0001-8460-8395
Burcu Balcik https://orcid.org/0000-0002-3575-1846

REFERENCES

Abdellaoui, M., L'Haridon, O., & Paraschiv, C. (2011). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science*, *57*(10), 1879–1895.

Abuoda, G., Hendrix, C., & Campo, S. (2021). Automatic tag recommendation for the UN Humanitarian Data Exchange. *BIRDS+ WEPIR@ CHIIR*, 4–10.

ACAPS. (2014). *Idleb–governorate profile–Syria needs analysis project*. https://reliefweb.int/report/syrian-arab-republic/idleb-governorate-profile-syria-needs-analysis-project-june-2014

Altay, N., & Labonte, M. (2014). Challenges in humanitarian information management and exchange: evidence from Haiti. *Disasters*, *38*(s1), 50–72.

Andres, J., Wolf, C. T., Cabrero Barros, S., Oduor, E., Nair, R., Kjærum, A., Tharsgaard, A. B., & Madsen, B. S. (2020). Scenario-based XAI for humanitarian aid forecasting. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–8). Association for Computing Machinery.

Azizi, S., Bozkir, C. D. C., Trapp, A. C., Kundakcioglu, O. E., & Kurbanzade, A. K. (2021). Aid allocation for camp-based and urban refugees with uncertain demand and replenishments. *Production and Operations Management*, *30*(12), 4455–4474.

Balcik, B., Beamon, B. M., Krejci, C. C., Muramatsu, K. M., & Ramirez, M. (2010). Coordination in humanitarian relief chains: Practices, challenges and opportunities. *International Journal of Production Economics*, *126*(1), 22–34.

Balcik, B., Silvestri, S., Rancourt, M.-È., & Laporte, G. (2019). Collaborative prepositioning network design for regional disaster response. *Production and Operations Management*, *28*(10), 2431–2455.

Bengio, Y., Lodi, A., & Prouvost, A. (2021). Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, *290*(2), 405–421.

Benini, A., Chataigner, P., Noumri, N., Parham, N., Sweeney, J., & Tax, L. (2017). *The use of expert judgment in humanitarian analysis—Theory, methods, applications*. Assessment Capacities Project - ACAPS.

Besiou, M., & Van Wassenhove, L. N. (2020). Humanitarian operations: A world of opportunity for relevant and impactful research. *Manufacturing & Service Operations Management*, *22*(1), 135–145.

Birge, J. R., & Louveaux, F. (2011). *Introduction to stochastic programming* (2nd ed.). Springer Series in Operations Research and Financial Engineering. Springer Science & Business Media.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer.

Chamola, V., Hassija, V., Gupta, S., Goyal, A., Guizani, M., & Sikdar, B. (2020). Disaster and pandemic management using machine learning: a survey. *IEEE Internet of Things Journal*, *8*(21), 16047–16071.

Comes, T., Van de Walle, B., & Van Wassenhove, L. (2020). The coordination-information bubble in humanitarian response: Theoretical foundations and empirical investigations. *Production and Operations Management*, *29*(11), 2484–2507.

Day, J. M., Melnyk, S. A., Larson, P. D., Davis, E. W., & Whybark, D. C. (2012). Humanitarian and disaster relief supply chains: A matter of life and death. *Journal of Supply Chain Management*, *48*(2), 21–36.

Dönmez, Z., Kara, B. Y., Karsu, Ö., & Saldanha-da Gama, F. (2021). Humanitarian facility location under uncertainty: Critical review and future prospects. *Omega*, *102*, 102393.

Dufour, É., Laporte, G., Paquette, J., & Rancourt, M.-È (2018). Logistics service network design for humanitarian response in East Africa. *Omega*, *74*, 1–14.

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

Ergun, Ö., Gui, L., Heier Stamm, J. L., Keskinocak, P., & Swann, J. (2014). Improving humanitarian operations through technology-enabled collaboration. *Production and Operations Management*, *23*(6), 1002–1014.

Farahani, R. Z., Lotfi, M., Baghaian, A., Ruiz, R., & Rezapour, S. (2020). Mass casualty management in disaster scene: A systematic review of OR&MS research in humanitarian operations. *European Journal of Operational Research*, *287*(3), 787–819.

Galindo, G., & Batta, R. (2013). Review of recent developments in OR/MS research in disaster operations management. *European Journal of Operational Research*, *230*(2), 201–211.

Gallien, J., Leung, N.-H. Z., & Yadav, P. (2021). Inventory policies for pharmaceutical distribution in Zambia: Improving availability and access equity. *Production and Operations Management*, *30*(12), 4501–4521.

Ghysels, E., Qian, Y., & Raymond, S. (2021). Ambiguity with machine learning: An application to portfolio choice. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3951595

Gralla, E., Goentzel, J., & Fine, C. (2016). Problem formulation and solution mechanisms: a behavioral study of humanitarian transportation planning. *Production and Operations Management*, *25*(1), 22–35.

Grass, E., & Fischer, K. (2016). Prepositioning of relief items under uncertainty: A classification of modeling and solution approaches for disaster management. *Logistics Management*, 189–202.

Gupta, S., Altay, N., & Luo, Z. (2019). Big data in humanitarian supply chain management: A review and further research directions. *Annals of Operations Research*, *283*(1), 1153–1173.

Gupta, S., Starr, M. K., Farahani, R. Z., & Matinrad, N. (2016). Disaster management from a POM perspective: Mapping a new domain. *Production and Operations Management*, *25*(10), 1611–1637.

Gutjahr, W. J., & Nolz, P. C. (2016). Multicriteria optimization in humanitarian aid. *European Journal of Operational Research*, *252*(2), 351–366.

Hammitt, J. K., & Shlyakhter, A. I. (2006). The expected value of information and the probability of surprise. *Risk Analysis*, *19*(1), 135–152.

Hewitt, M., Ortmann, J., & Rei, W. (2022). Decision-based scenario clustering for decision-making under uncertainty. *Annals of Operations Research*, *315*(2), 747–771.

HNO. (2019). Humanitarian Needs Overview 2019: Syrian Arab Republic. https://hno-syria.org/data/downloads/en/full_hno_2019.pdf

Hoffman, F. O., & Hammonds, J. S. (1994). Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis*, *14*(5), 707–712.

Hosseinnezhad, D., & Saidi-mehrabad, M. (2018). Data fusion and information transparency in disaster chain. *International Journal of Innovation, Management and Technology*, *9*(4), 152–159.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441–458.

Kılcı, F., Kara, B. Y., & Bozkaya, B. (2015). Locating temporary shelter areas after an earthquake: A case for Turkey. *European Journal of Operational Research*, *243*(1), 323–332.

Kınay, Ö. B., Kara, B. Y., Saldanha-da Gama, F., & Correia, I. (2018). Modeling the shelter site location problem using chance constraints: A case study for Istanbul. *European Journal of Operational Research*, *270*(1), 132–145.

King, A. J., & Wallace, S. W. (2012). *Modeling with stochastic programming*. Springer Series in Operations Research and Financial Engineering. Springer.

Liberatore, F., Pizarro, C., Blas, C. S., Ortuno, M. T., & Vitoriano, B. (2013). Uncertainty in humanitarian logistics for disaster management: A review. In Vitoriano, B., Montero, J., & Ruan, D. (Eds.), *Decision aid models for disaster management and emergencies* (pp. 45–74). Atlantis Press.

Linardos, V., Drakaki, M., Tzionas, P., & Karnavas, Y. L. (2022). Machine learning in disaster management: Recent developments in methods and applications. *Machine Learning and Knowledge Extraction*, *4*(2), 446–473.

Lorca, Á., Çelik, M., Ergun, Ö., & Keskinocak, P. (2017). An optimization-based decision-support tool for postdisaster debris operations. *Production and Operations Management*, *26*(6), 1076–1091.

Marić, J., Galera-Zarco, C., & Opazo-Basáez, M. (2022). The emergent role of digital technologies in the context of humanitarian supply chains: A systematic literature review. *Annals of Operations Research*, *319*(1), 1003–1044.

Mazyavkina, N., Sviridov, S., Ivanov, S., & Burnaev, E. (2021). Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, *134*, 105400.

McCoy, J. H., & Lee, H. L. (2014). Using fairness models to improve equity in health delivery fleet management. *Production and Operations Management*, *23*(6), 965–977.

Ni, W., Shu, J., & Song, M. (2018). 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*, *27*(1), 160–183.

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

O'Brien, S. (2017). This is how we build a stronger, data-driven humanitarian sector. https://www.weforum.org/agenda/2017/01/this-is-how-we-build-a-stronger-data-driver-humanitarian%20sector/

OCHA. (2018). Services humanitarian response: IDP sites integrated monitoring matrix (ISIMM). https://www.humanitarianresponse.info/en/operations/stima/camp-coordination-management/documents

Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., Briant, J., Millet, P., Reinhard, F., Parkan, M., & Joost, S. (2016). Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big Data*, *4*(1), 47–59.

Ozbay, E., Çavuş, Ö., & Kara, B. Y. (2019). Shelter site location under multi-hazard scenarios. *Computers & Operations Research*, *106*, 102–118.

Paulus, D., Meesters, K., & Van de Walle, B. A. (2018). Turning data into action: supporting humanitarian field workers with open data. In K. Boersma & B. Tomaszewski (Eds.), *Proceedings of the 15th ISCRAM conference* (Vol. 15, pp. 1030–1040). ISCRAM.

Rahimian, H., & Mehrotra, S. (2019). *Distributionally robust optimization: A review*. arXiv. https://doi.org/10.48550/arXiv.1908.05659

Raymond, N., & Al Achkar, Z. (2016). *Data preparedness: connecting data, decision-making and humanitarian response*. Harvard Humanitarian Initiative. https://reliefweb.int/sites/reliefweb.int/files/resources/data.pdf

REACH. (2018). *Situation overview: Idleb governorate and surrounding areas*. https://reliefweb.int/sites/reliefweb.int/files/resources/reach_syr_situation_overview_idleb_governorate_and_surrounding_areas_needs_assessment_may_2018.pdf

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computer Application Math*, *20*, 53–65.

Ruesch, L., Tarakci, M., Besiou, M., & Van Quaquebeke, N. (2022). Orchestrating coordination among humanitarian organizations. *Production and Operations Management*, *31*(5), 1977–1996.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Shi, P., Helm, J. E., Chen, C., Lim, J., Parker, R. P., Tinsley, T., & Cecil, J. (2023). OperatioEShns (management) warp speed: Rapid deployment of hospital-focused predictive/prescriptive analytics for the Covid-19 pandemic. Production and Operations Management, *32*(5), 1433–1452.

Snow, A. (2010). Ambiguity and the value of information. *Journal of Risk and Uncertainty*, *40*(2), 133–145.

Sokat, K. Y., Zhou, R., Dolinskaya, I. S., Smilowitz, K., & Chan, J. (2016). Capturing real-time data in disaster response logistics. *Journal of Operations and Supply Chain Management*, *9*(1), 23–54.

Starr, M. K., & Van Wassenhove, L. N. (2014). Introduction to the special issue on humanitarian operations and crisis management. *Production and Operations Management*, *23*(6), 925–937.

Stauffer, J. M., Pedraza-Martinez, A. J., & Van Wassenhove, L. N. (2016). Temporary hubs for the global vehicle supply chain in humanitarian operations. *Production and Operations Management*, *25*(2), 192–209.

Sun, W., Bocchini, P., & Davison, B. D. (2020). Applications of artificial intelligence for disaster management. *Natural Hazards*, *103*(3), 2631–2689.

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

Swamy, V., Chen, E., Vankayalapati, A., Aggarwal, A., Liu, C., Mandava, V., & Johnson, S. (2019). Machine learning for humanitarian data: Tag prediction using the HXL standard. In *KDD '19: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. Workshop on Data for Social Impact, August 04–08, 2019, Anchorage, AK. p. 3. ACM, New York, NY, USA.

Taylor, K., Zarb, S., & Jeschke, N. (2021). Ambiguity, uncertainty and implementation. *International Review of Public Policy*, *3*(1), 1638. https://journals.openedition.org/irpp/1638

UN Refugee Agency. (2021). *Syria refugee crisis*. https://www.unrefugees.org/emergencies/syria/

Van de Walle, B., & Comes, T. (2015). On the nature of information management in complex and natural disasters. *Procedia Engineering*, *107*, 403–411.

Van Wassenhove, L. N., & Besiou, M. (2013). Complex problems with multiple stakeholders: How to bridge the gap between reality and OR/MS? *Journal of Business Economics*, *83*(1), 87–97.

Vesselinova, N., Steinert, R., Perez-Ramirez, D. F., & Boman, M. (2020). Learning combinatorial optimization on graphs: A survey with applications to networking. *IEEE Access*, *8*, 120388–120416.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.

Yáñez-Sandivari, L., Cortés, C. E., & Rey, P. A. (2020). Humanitarian logistics and emergencies management: New perspectives to a sociotechnical problem and its optimization approach management. *International Journal of Disaster Risk Reduction*, *52*, 101952.

Yoo, E., Rabinovich, E., & Gu, B. (2020). The growth of follower networks on social media platforms for humanitarian operations. *Production and Operations Management*, *29*(12), 2696–2715.

Zagorecki, A. T., Johnson, D. E., & Ristvej, J. (2013). Data mining and machine learning in the context of disaster and crisis management. *International Journal of Emergency Management*, *9*(4), 351–365.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Grass, E., Ortmann, J., Balcik, B., & Rei, W. (2023). A machine learning approach to deal with ambiguity in the humanitarian decision-making. *Production and Operations Management*, *32,* 2956–2974. https://doi.org/10.1111/poms.14018