Annals of Operations Research
https://doi.org/10.1007/s10479-022-04990-z

ORIGINAL RESEARCH

Robust weighted aggregation of expert opinions in futures studies

Marco Marozzi[1] · **Mario Bolzan**[2] · **Simone Di Zio**[3]

Accepted: 12 September 2022

Abstract
Expert judgments are widespread in many fields, and the way in which they are collected and the procedure by which they are aggregated are considered crucial steps. From a statistical perspective, expert judgments are subjective data and must be gathered and treated as carefully and scientifically as possible. In the elicitation phase, a multitude of experts is preferable to a single expert, and techniques based on anonymity and iterations, such as Delphi, offer many advantages in terms of reducing distortions, which are mainly related to cognitive biases. There are two approaches to the aggregation of the judgments given by a panel of experts, referred to as *behavioural* (implying an interaction between the experts) and *mathematical* (involving non-interacting participants and the aggregation of the judgments using a mathematical formula). Both have advantages and disadvantages, and with the mathematical approach, the main problem concerns the subjective choice of an appropriate formula for both normalization and aggregation. We propose a new method for aggregating and processing subjective data collected using the Delphi method, with the aim of obtaining robust rankings of the outputs. This method makes it possible to normalize and aggregate the opinions of a panel of experts, while modelling different sources of uncertainty. We use an uncertainty analysis approach that allows the contemporaneous use of different aggregation and normalization functions, so that the result does not depend on the choice of a specific mathematical formula, thereby solving the problem of choice. Furthermore, we can also model the uncertainty related to the weighting system, which reflects the different expertise of the participants as well as expert opinion accuracy. By combining the Delphi method with the robust ranking procedure, we offer a new protocol covering the elicitation, the aggregation and the processing of subjective data used in the construction of Delphi-based future scenarios. The method is very flexible and can be applied to the aggregation and processing of any subjective judgments, i.e. also those outside the context of futures studies. Finally, we show the validity, reproducibility and potential of the method through its application with regard to the future of Italian families.

✉ Simone Di Zio
s.dizio@unich.it

1 Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy

2 Department of Statistical Sciences, University of Padua, Padua, Italy

3 Department of Legal and Social Sciences, University "G. d'Annunzio" of Chieti-Pescara, Chieti, Italy

Published online: 27 September 2022

Keywords Monte Carlo methods · Robust ranking · Uncertainty analysis · Delphi-based scenarios · Expertise

1 Introduction

In most contexts, existing data are not considered sufficient to implement effective policies and/or to make the best management decisions. In these cases, decision makers make use of other sources of information, among which the judgments of experts are widely used (Colson & Cooke, 2018). Experts have become indispensable in organizations because they "fill gaps in data and in the understanding of existing or missing data" (Benini et al., 2017, p. 1). From marketing (Larréché & Moinpour, 1983) to project management (Szwed, 2016), medicine (Bojke et al., 2021) to the economy (Usher & Strachan, 2013), and natural resource management (Hemming et al., 2018) to risk assessment (Hanea et al., 2021) and terrorism (Gordon et al., 2015), the use of experts and their structured judgments has become widespread.

Whether the subjective opinions coming from expert evaluations have scientific value or not is still under debate, but all authors agree that the judgments of a group are better than those of a single expert. Furthermore, both the way in which the judgments are collected and the procedure by which they are aggregated to form a single evaluation are considered crucial (O'Hagan, 2019). If within a properly built panel of experts, judgments are collected and aggregated according to validated protocols, then the subjective data have scientific validity, just like any other type of data (O'Hagan, 2019).

Expert judgments are widely used in various fields, and when dealing with future scenarios (and more generally in futures studies), since data do not exist with regard to the future, the use of experts is essential. In many futures studies (FS) and strategic foresight methodologies, experts are used extensively, and the main problem remains that of using a formalized and structured procedure for first collecting and then combining experts' judgments.

Statistically speaking, experts' judgments are subjective data and should be gathered and treated as carefully and scientifically as possible (O'Hagan, 2019). For collecting subjective data, informally asking a single expert, interviewing several experts or arranging a focus group are useful techniques, but it is widely recognized that numerous cognitive biases invalidate the judgments made by experts during these procedures (see, among others, Bonaccorsi et al., 2020 and O'Hagan, 2019). The following is a short list of the most common distortions that occur during face-to-face meetings: 1) *Leadership*: when the highest-ranking person expresses an opinion, the others usually tend to follow him/her, and the risk is that many participants do not feel free to express their opinions for fear of conflict with the leader. 2) *The spiral of silence*: those whose opinions are in line with the majority feel more confident in expressing their views, while others fear that sharing their opinions could cause social ostracism. Therefore, the latter remain silent, and this leads to a spiralling process, in which the opinions of the minority are more and more marginalized and suppressed (Di Zio & Staniscia, 2014). 3) *Groupthink*: the pressure to conform within a group interferes with the correct analysis of the problem and produces poor group decisions. In other words, groupthink involves distortion to minimize conflicts and to avoid challenging others' views, during which individuals withhold their personal opinions (McCauley, 1998). 4) *Group polarization*: a distortion that affects people who work in groups and pushes them towards accepting opinions that are more extreme than their own individual beliefs (Isenberg, 1986).

To capture expert knowledge as objectively as possible, the gathering process must be structured carefully to avoid (or at least to minimize) such biases (O'Hagan, 2019). In the relevant literature, a number of procedures—actual *elicitation protocols*—have been developed that attempt to mitigate the distortions that are triggered when the judgments are collected and which can compromise the validity of the subjective estimates.

The technical activity that leads an expert to form and express an opinion is called *expert elicitation* (O'Hagan, 2019). Apart from the elicitation situation (e.g. interactive expert groups or Delphi-like procedures) and the mode of collecting the judgments (e.g. telephone or computer assisted), there are two crucial phases to any expert elicitation: a) the elicitation technique and b) the aggregation of views. A multiplicity of experts is preferable to a single expert because it guarantees the diversity of knowledge and background but involves the problem of aggregating the different points of view into a single solution, which implies the combination of several pieces of data (Meyer & Booker, 1991). The relevant literature distinguishes two main approaches of aggregation, known as *behavioural*—characterized by interactions among experts that seek consensus—and *mathematical*, in which separate judgments are obtained from non-interacting experts and then combined using a mathematical formula (Clemen & Winkler, 1999; Colson & Cooke, 2018; Meyer & Booker, 1991).

Many authors recommend the validation of expert judgments (Cooke, 1991; Colson & Cooke, 2018), which means, on the one hand, that the judgments should reflect the beliefs of the experts (and therefore be sheltered from cognitive biases) and, on the other, that these beliefs should reflect reality. Since the former cannot be measured, validation consists of a comparison between the elicited judgments and the observed data, but this is only possible if observed data exist (Colson & Cooke, 2018). In the case of future scenarios and, more generally, in FS, the crucial point is that no data exist with regard to the future. The French philosopher de Jouvenel (1967) points out the difference between accomplished facts, which have taken a form that is no longer modifiable (which he calls *facta*), and what is instead in the making and can still be realized in different forms (called, by contrast, *futura*). Therefore, the main problem with the use of expert opinions in FS, like in many other fields, remains that of collecting and aggregating the judgments.

The Delphi method is one of the most widely used techniques for gathering subjective data from groups of experts (Dalkey & Helmer, 1963; Linstone & Turoff, 1975). In the context of FS, the Delphi method is very popular and is often used in combination with the scenario method. A future scenario is a description of a possible future situation with the paths of development leading to that future (Kosow & Gaßner, 2008). When the outputs of Delphi are used as inputs for scenario building, so-called *Delphi-based scenarios* provide the context (von der Gracht & Darkow, 2010; Di Zio et al., 2021), and this paper will use this approach. While the Delphi technique is considered one of the most robust methods for the elicitation of subjective data, when moving from the Delphi outputs to the scenarios, the crucial steps are precisely those of the aggregation of expert judgments and the statistical data processing necessary for the subsequent steps of the scenario development.

In short, subjective data from experts are important and sometimes needful, but if not collected, normalized, aggregated and treated correctly, they can boomerang on the goals and the results of the research. In the words of Ayyub (2001), they can be a double-edged weapon:

> "Experts, despite their importance and value, can be double-edged swords. They can make valuable contributions from their deep base of knowledge, but those contributions may also contain their own biases and pet theories. Therefore, selecting experts,

eliciting their opinions, and aggregating their opinions must be performed and handled carefully, with full recognition of the uncertainties inherent in those opinions."

In this paper, we propose a new method for aggregating and processing subjective data collected with Delphi with the aim of obtaining robust rankings of Delphi projections that are useful for the subsequent phase of future scenario development. As we will see, the method allows the aggregation and normalization of the opinions of a panel of experts while modelling the various sources of uncertainty. Our proposal is based on uncertainty analysis, which implies the simultaneous use of different aggregation and normalization functions, so the result does not depend on the choice of a particular mathematical formula. This allows us to state that the method we propose in this paper preserves many of the advantages of both the mathematical and behavioural approaches, while limiting their disadvantages. The method allows all the opinions expressed by experts in the different rounds to be taken into account, but since there is no face-to-face communication, it avoids many cognitive biases (leadership, the spiral of silence, groupthink and group polarization), and in the aggregation phase, it does not require the subjective choice of a specific formula. Furthermore, with uncertainty analysis, we can also model the uncertainty related to the system of weights reflecting the different expertise of the participants as well as the expert opinion accuracy (with accuracy being defined as both under/overestimation and the precision of the expert scores). In general, it is shown that the method is very flexible because a researcher can easily incorporate his/her preferences on normalizing and aggregating data, as well as on modelling weights and score accuracy. The results of the method can be represented graphically in a very clear manner through a succession of intervals for each item. The information provided, for example, with regard to more or less overlapping intervals, as well as the lengths of the same intervals, can then be used for subsequent analyses.

A limitation, rather than a disadvantage, of the method is that in its current form it can only be applied to the ranking of the items but not to their ratings. The method is also conceptually applicable to the ratings, but at the price of reduced flexibility, because the ratings are not always comparable when the formula adopted to combine the opinions of the experts varies, unlike the rankings, which are always comparable since they are always ordinal numbers. An open problem is the full integration of the various qualitative phases (first focus groups and then the Delphi steps) within the proposed method. In fact, in this paper, while we also illustrate how it may be possible to integrate the initial phases of Delphi as well, only the last phase of Delphi provides input for our method.

By applying the method to the future of Italian families, we will show how the resulting rankings can be used as a base to build futures scenarios. By combining the Delphi method with this new robust ranking procedure, we offer a new protocol covering the elicitation, aggregation and processing of subjective data used in the construction of future scenarios. However, the method is very flexible and can be applied to the aggregation and processing of any kind of subjective judgments, even those outside the FS context.

The remainder of this paper is organized as follows: Sect. 2 contains the literature review, while Sect. 3 explains the proposed methodology. Section 4 contains a description of how the method was applied to future scenarios of Italian families along with the main results obtained. Finally, Sect. 5 contains the conclusion, limitations and proposals for future developments.

2 Technical literature review

FS exploit a wide range of methods, including some borrowed and adapted from other disciplines and others specifically designed for studying the future. It is not possible to review all the methods here, so we refer to the specialized literature (see Glenn & Gordon, 2009). However, among the best known and most widely used methods born in the very years when the future was being studied with greater scientific rigor are the Delphi and scenario techniques.

The Delphi method was developed in the 1950s at the RAND Corporation by Olaf Helmer and Norman Dalkey and then made public by Theodore J. Gordon and Olaf Helmer in 1964 (Gordon & Helmer, 1964). It is one of the most widely used and accepted techniques for gathering information from experts (Dalkey & Helmer, 1963; Linstone & Turoff, 1975), and over the course of more than half a century, since its inception, it has seen thousands of applications and numerous methodological variants. A selected panel of experts answers a series of questionnaires, containing both the research questions and feedback from prior questionnaires. Key features of the Delphi method are anonymity, iteration, controlled feedback and the statistical aggregation of responses (Linstone & Turoff, 1975; Rowe & Wright, 1999), and one of the most frequent objectives (but not necessarily the only one) is consensus on the issues under investigation. It is worth noting that a Delphi panel involves non-probability sampling techniques (e.g. purposive sampling or criterion sampling).

Although, as mentioned, there are many variations of this method, the main steps of classical Delphi are as follows: (1) the selection of the panel of experts; (2) the construction and submission of the first questionnaire; (3) the aggregation of the first-round responses, typically with appropriate statistical synthesis; (4) the submission of the second questionnaire: experts are asked to reassess their judgments regarding the same questions by considering the provided statistical synthesis; (5) the aggregation of the second-round responses; (6) the submission of the third questionnaire with the inclusion of the comments from the previous round; and (7) the iteration of steps 3–6 until a stopping criterion is reached. From the second round onwards, each expert has the chance to revise his/her evaluations, and this produces, at least in principle, a gradual reduction of the variability in the distribution of the judgments, thus triggering consensus.

Delphi logic makes it possible to overcome and/or minimize the distortions typical of face-to-face communication techniques, which are basically cognitive biases (Bonaccorsi et al., 2020; O'Hagan, 2019). Starting from the pioneering research of Tversky and Kahneman (1974), psychologists identified many biases in human judgments, and experts are not exempt from such heuristics. With the isolation and anonymity of participants and non-synchronous communication, Delphi resolves many face-to-face distortions (Di Zio et al., 2021).

Many studies demonstrate how Delphi beats many other elicitation techniques (Rowe & Wright, 1999), and in the context of FS, it is often used in combination with the scenario method. A future scenario can be defined as a description of a future situation together with the paths of development leading to it. It is a hypothetical construct and not a description of the future; therefore, the aim is not to predict the future (forecast) but rather to highlight the crucial projections of possible futures (foresight) and the key variables that will drive these developments (Kosow & Gaßner, 2008; Schoemaker, 1995). Like Delphi, the scenario method originated during the 1950s in a military context but is now commonly used for long-term planning. The goal of scenario planning is to develop a number (generally 3–4) of alternative future scenarios that taken together should cover, to a certain extent, possible, plausible, probable and surprising futures (Bishop et al., 2007; Fritschy & Spinler, 2019).

Future scenarios are useful in supporting decision makers with regard to unveiling uncertainties and/or potential future threats (Bishop et al., 2007; Schoemaker, 1995).

The two methods, Delphi and scenario, are often combined, as many studies have shown that the outputs of scenario analysis can be used as inputs to improve a Delphi study, and the results of Delphi can also be used to facilitate scenario development (Nowack et al., 2011). The latter approach is also known as *Delphi-based scenarios* and is the most widespread. In fact, a future scenario is based on future projections, and Delphi is especially useful for the analysis of future projections (von der Gracht & Darkow, 2010). In a Delphi-based scenario procedure, a crucial phase involves the correct use of the Delphi results for the development of the scenarios. The judgments of the panel of experts resulting from the last Delphi round must be aggregated carefully to create a unique judgment for each projection, and then the aggregated results must be properly grouped to create clusters that form the basis for the development of the future scenarios (Di Zio et al., 2021).

The aggregation of the opinions of the experts who make up the panel is a problem that concerns not only Delphi-based scenarios but also all other contexts in which expert opinions are elicited. There are two main approaches to aggregation: *mathematical* and *behavioural* (Clemen & Winkler, 1999; Colson & Cooke, 2018; Meyer & Booker, 1991; O'Hagan, 2019). In the mathematical approach, single judgments are elicited from each expert on the panel and subsequently combined in a unique final solution using a mathematical formula, also called the *pooling rule*. In the behavioural approach, the experts of the panel are asked to discuss their opinions and then to come up with a unique final solution, which represents the consensus of the whole group.

Both approaches have advantages and disadvantages (O'Hagan, 2019). On the one hand, the mathematical approach solves the problems derived from the cognitive biases arising from face-to-face interaction (leadership, the spiral of silence, groupthink and group polarization), but, since there are many pooling rules, and no single one is considered to be the best, the researcher must make a subjective choice. Furthermore, mathematical aggregation does not highlight specific or particular opinions or the motivations of those who disagree with the majority, so the final result may not be an expression of any of the experts' thoughts (Benini et al., 2017). On the other hand, the behavioural approach does not imply subjectively choosing an aggregation formula and allows everyone's opinions to be taken into account. However, it involves many problems in seeking consensus among the experts and also includes the inevitable group cognitive biases. Moreover, the behavioural approach is often time consuming because generally conflicting points of view are difficult to agree on (Benini et al., 2017).

Because in the Delphi method there is no interaction between experts, and the result is the product of mathematical aggregation, it can be considered as part of the mathematical family. However, it does involve seeking consensus and entails a particular form of interaction between the participants, not face to face but through a supervised arrangement. This interaction occurs starting from the second round, when each expert receives, in the form of statistical summaries, the results of the previous consultation. From the second round onwards, the experts can provide anonymous reasons for their judgments to which, starting from the third round (and always anonymously), the others can respond. In this way, a kind of anonymous and remote debate is triggered, also called an *anonymous conference* (Di Zio & Pacinelli, 2019), which therefore implies a certain form of interaction. Undoubtedly, it is an interaction that, unlike face-to-face methods, eliminates all the cognitive biases typical of meetings in which people are in the same place and discussing issues face to face with limited time to finish the work. These are the reasons why some authors classify Delphi as

a mathematical method (O'Hagan, 2019) and others consider it to be a behavioural method (Benini et al., 2017).

According to the aforementioned observations, the method proposed in this paper (based on the analysis of Delphi results) exploits the advantages of both approaches (*mathematical* and *behavioural*), and it can be classified as a *mixed method*, that is, an approach that uses and integrates multiple methods and, above all, mixes qualitative and quantitative methodologies. In studying complex phenomena, many scholars claim that a mixed-method approach is desirable, given the need to analyse the problem from many perspectives (Johnson & Onwuegbuzie, 2004; Sale et al., 2002).

Finally, an open question in the literature concerns the way in which the competence or degree of expertise of the participants can be measured and, consequently, how experts, and/or their evaluations, can be weighed (Sossa et al., 2019). Expert weighting is a continuous source of uncertainty in the aggregation of assessments, and in the methodology proposed here, we will also address the issue of how expert weights can be modelled. Although expert weights affect the results of any Delphi study, the problem of modelling them has received limited attention in the literature.

3 Aggregating expert opinions by combining mathematical and behavioural approaches

3.1 The proposed approach

Let X_{ji} denote the assessment of item $i = 1, \ldots, I$ according to expert $j = 1, \ldots, J$. Examples of common assessments are the probability of occurrence, impact, plausibility, relevance and evolution. Generally, X_{ji} is a (numeric) score. Sometimes, X_{ji} is an ordered categorical variable. In this subsection, we describe a very general and flexible approach to combining expert opinions, which can easily be customized to handle the case of ordered categorical variables.

The most familiar way to combine expert opinions is by averaging X_{ji} over the experts using the arithmetic mean (Szwed, 2016; Cooke, 1991). This approach corresponds to.

$$_0C_i = \frac{1}{J}\sum_{j=1}^{J} X_{ji}.$$

The arithmetic mean is very simple, and this is its main advantage because it is easily comprehensible to all stakeholders. However, looking at the arithmetic mean from a different perspective, its simplicity is also its main disadvantage. In fact, the arithmetic mean is very limited because.

(i) the experts have the same weights; and
(ii) the item assessments X_{ji}, $j = 1, \ldots, J$ are not normalized or standardized.

Point (i) is a drawback because it is generally advisable or desirable to weight experts differently according to their expertise, usually by self-weighing, or, more generally, with performance weights resulting from a validation procedure, when possible (Cooke, 1991). Point (ii) is also a drawback because, in general, the assessment vectors of experts $(X_{j1}, \ldots, X_{jI})$ $j = 1, \ldots, J$ are not comparable due to their different locations and variability. Therefore, we do not suggest combining the raw assessments but rather adjusting them to achieve comparability. These considerations lead to.

${}_{(.)1}C_i = \sum_{j=1}^{J} w_j N(X_{ji}) / \sum_{j=1}^{J} w_j$,where w_j denotes the weight of expert j, and $N(.)$ denotes a normalization function. The usual constraints apply to the weights: $w_j > 0$ and $\sum_{j=1}^{J} w_j = 1$.

The most familiar normalization functions are

$$N_1(X_{ji}) = \frac{X_{ji} - min(X_{j.})}{max(X_{j.}) - min(X_{j.})}$$

and.

$N_2(X_{ji}) = \frac{X_{ji} - mean(X_{j.})}{sd(X_{j.})}$,where $X_{j.} = (X_{j1}, ..., X_{jI})$ is the vector of expert j's assessments, $mean(X_{j.}) = \frac{1}{I}\sum_{i=1}^{I} X_{ji}$ and $sd(X_{j.}) = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(X_{ji} - mean(X_{j.}))^2}$. N_1 is the linear scaling in the min–max range, and N_2 is the z score standardizing $X_{j.}$ values such that the adjusted values have a 0 mean and 1 standard deviation.

Other quite familiar normalization functions are.

$N_3(X_{ji}) = \frac{X_{ji} - median(X_{j.})}{mad(X_{j.})}$,where $mad(X_{j.}) = median(|X_{j.} - median(X_{j.})|)$;

$N_4(X_{ji}) = \sum_{h=1}^{I}(X_{ji}X_{jh})$,where $(X_{ji}X_{jh}) = 1$ if $X_{ji}X_{jh}$ and 0 otherwise; and

$$N_5(X_{ji}) = \frac{X_{ji}}{\sqrt{\sum_{i=1}^{I} X_{ji}^2}}$$

and.

$N_6(X_{ji}) = \frac{X_{ji}}{max(X_{j.})}$.

N_3 is a robust version of N_2 and is preferred to N_2 when analysing heavy-tailed or highly-skewed data. N_4 is the rank transformation and is another option for a robust normalization function. N_5 is an example of vector normalization using the Euclidean norm. N_6 is a linear scaling similar to N_1. N_6 is preferred to N_1 when 0 adjusted values are not desirable because they can lead to computation issues in the combination step of the procedure. However, it is also possible to consider.

$N_7(X_{ji}) = \frac{X_{ji} - min(X_{j.}) + \frac{1}{I}}{max(X_{j.}) - min(X_{j.}) + \frac{2}{I}}$,where $1/I$ and $2/I$ are added to the numerator and denominator, respectively, to avoid 0 values but retain linear scaling in the min–max range.

The aggregation of expert opinions using the arithmetic mean is equivalent to the additive rule of combination. There are other types of combination. A quite familiar one is the multiplicative rule corresponding to the geometric mean:

${}_{(.)2}C_i = \prod_{j=1}^{J} N(X_{ji})^{W_j}$.

Note that both ${}_1C$ and ${}_2C$ are special case of the generalized power mean $\left[\sum_{j=1}^{J} N(X_{ji})^p w_j\right]^{1/p}$ for $p = 1$ and $p \to 0$, respectively. Other particular cases are

$${}_{(.)3}C_i = min[N(X_{1i}), \ldots, N(X_{Ji})]$$

and

$${}_{(.)4}C_i = max[N(X_{1i}), \ldots, N(X_{Ji})]$$

for $p \to$ -∞ and $p \to \infty$, respectively. ${}_4C$ is also called the Tippett combination. Other combination functions are.

$${}_{(.)5}C_i = -\sum_{j=1}^{J} log\big(1 - N(X_{ji})\big)w_j,$$

$${}_{(.)6}C_i = \sum_{j=1}^{J} log\left(\frac{N(X_{ji})}{1 - N(X_{ji})}\right)w_j$$

and.

${}_{(.)7}C_i = \sum_{j=1}^{J} F^{-1}\big(N(X_{ji})\big)w_j$,where Φ^{-1} is the quantile function of the standard normal distribution. ${}_5C$ is the Fisher combining function, ${}_6C$ is the logistic combining function and ${}_7C$ is the Liptak combining function. ${}_6C$ and ${}_7C$ are particular cases of $\sum_{j=1}^{J} F^{-1}\big(N(X_{ji})\big)w_j$, where F^{-1} is the quantile function of a continuous random variable. It is important to note that the ${}_4C$ to ${}_7C$ combining functions are quite familiar within the framework of the non-parametric combination of dependent tests (see Pesarin & Salmaso, 2010).

It is important to emphasize that there is no best combining function as there is no best normalization function. Every function has advantages and disadvantages. As we have already discussed, N_3 and N_4 are robust against heavy-tailed and highly-skewed data, so when outliers are present. However, they have disadvantages too: N_3 should not be used when the data have low variability because both the median and the median absolute deviation from the median are unstable. N_4 assesses the order of the data and not the data values, leading to a loss of information that can be relevant.

We turn our attention to combining functions. The additive rule is fully compensatory because low scores are compensated by higher scores. In some fields this is preferred, an example being the labour market, where compensating differentials are typical, and workers are often offered different combinations of salary and working conditions and can decide to work very close to home in exchange for a lower salary. However, there are fields where partial compensatory combinations, like geometrics, are preferred; an example is the study of economic development, where high industrialization should not fully compensate for a high level of pollution or land use. Another important aspect when selecting the combining function is the compatibility with the normalized data. The additive combination has the advantage of being compatible with all the normalization functions. The multiplicative rule cannot be used with negative or zero values; therefore, it is not suitable for z scores N_2 and N_3, for example. The Fisher, logistic and Liptak rules are only compatible with linear scaling N_7. It is worth noting that some of these aggregation rules, which involve direct sums, are only appropriate if all the expert opinions are positively correlated; otherwise, there may be compensation issues. However, when the method is applied, like in our application, in a Delphi context, this assumption is not strict because the aggregation method only intervenes after the Delphi rounds have concluded. This means that if the convergence of opinions has been triggered during the Delphi rounds, in the end, the experts will reach a consensus with similar values.

The problem of selecting normalization and combination functions can be addressed by borrowing uncertainty analysis from the vast literature on composite indicators. Uncertainty analysis is a Monte Carlo-based technique that allows a very flexible modelling of normalization and combination (see Saisana et al., 2005 and Marozzi, 2021). The rationale behind uncertainty analysis is based on modelling the sources of uncertainty in aggregating expert

opinions. There is always a certain degree of subjectivity in selecting a particular normalization and combination function, as there is in assigning weights to experts. Another important source of uncertainty is related to expert opinion accuracy. If a best formula existed to aggregate expert opinions, then its choice would be objective. Since there is no such formula, the choice of how to combine opinions becomes subjective. Uncertainty analysis simultaneously considers different formulas to aggregate expert opinions by varying the normalization and combination functions as well as the expert weights and opinion accuracy in each iteration of the Monte Carlo procedure. Therefore, the results do not depend on the particular formula subjectively selected by the scholar; instead, they are based on a very large number of different formulas with varying normalization and combination functions and expert weights and opinion accuracy. The uncertainty algorithms used here are explained in detail in Subsections 4.2 and 4.3.

Another advantage of uncertainty analysis is that it is possible to assess the effect of the sources of uncertainty on the aggregation of expert opinions, allowing the assessment of whether the results are robust or volatile with respect to the sources of uncertainty. This is a very important point because the strength of the message that is transmitted through the data analysis depends on the robustness of the results. Let M be the number of Monte Carlo iterations. M is set at a large number like 10,000 to account for continuous sources of uncertainty (expert weights and expert opinion accuracy). Note that normalization and combinations are discrete sources of uncertainty. The output of uncertainty analysis is an $M \times I$ matrix $Q = [q_{mi}]$, whose m-th row $(q_{m1}, \ldots, q_{mI})$ contains the combined scores (after normalization) of the I items corresponding to the m-th iteration of uncertainty analysis. We would like to rank the items according to the expert assessments. Therefore, we compute the ranks of the output matrix row-wise. The resulting matrix is denoted by $R = [r_{mi}]$. R contains item rankings, the rows, each of which corresponds to a particular combination of normalized scores as well as expert weights and expert opinion accuracy (as shown in Subsections 3.2 and 3.3). Column i $(r_{1i}, \ldots, r_{Mi})$ of R contains the uncertainty distribution of the rank of item i, reflecting a plurality of different formulas aggregating expert opinions. Note that using the traditional approach, one obtains just one rank for item i, making it impossible to assess the uncertainty and then the robustness of the result. Moreover, a unique rank value is dependent on the particular formula used to compute the combined scores. A different formula could lead to a different value. Instead, the uncertainty distribution of rank i can be summarized by computing the median, which is almost unaffected by the way expert opinions are aggregated (see Di Zio et al., 2021). The robustness of the results can be addressed by computing the 5^{th}–95^{th} percentile uncertainty intervals. Short intervals mean that the corresponding rank value is very stable across different aggregating formulas. Conversely, long intervals mean that the result depends on the selection of particular formulas. Markedly overlapping intervals among several item rankings mean that those ranks are rather similar.

3.2 Modelling the weights of experts

How can experts be weighted? This is an open question. Expert weighting is an important source of uncertainty when aggregating expert opinions. It is common to link expert weight to expertise, assigning different weights to experts according to their levels of expertise. There is no agreement among scholars on how to measure expertise (Sossa et al., 2019). Common methods are expertise evaluation by peers, the number of publications or citations, citation impact and the number of years spent working on the subject. Another very common situation

is when experts are requested to self-evaluate their level of expertise related to each of the I items, and this leads to a vector of weights for each expert $(w_{j1}, \ldots, w_{jI})$. The vector has a length of less than I in the case that experts are requested to self-evaluate their expertise for groups of items, where different groups relate to different dimensions of the phenomenon.

A simple approach is to assign each expert a constant weight, as a summary of $(w_{j1}, \ldots, w_{jI})$, using the mean, for example:

$$w_j = \frac{1}{J} \sum_{j=1}^{J} w_{ji}.$$

The mean can be preferred to using $(w_{j1}, \ldots, w_{jI})$ directly if it is expected or assumed that experts are overly lenient in self-evaluating their expertise with regard to some items while being overly strict in self-evaluating their expertise in terms of other items, because the mean compensates lower values with higher ones. On the other hand, one may prefer to directly use $w_{j1}, \ldots, w_{jI}$ values. Uncertainty related to expert weights can be modelled similarly to normalization and combination using uncertainty analysis. It is generally assumed that w_{ji} values, being estimates, are affected by measurement error. Uncertainty can be modelled by applying an additive random error. In this case, we compute it in each step of the Monte Carlo procedure.

${}_1u\left(w_{ji}\right) = w_{ji} + \eta_{ji}, \forall j, i$,where η_{ji} is a random error term. For maximum flexibility, the error term can always be the same $\eta_{ji} = \eta$ or be different item-wise $\eta_{ji} = \eta_i$ if different precision is assumed in self-evaluating one's expertise regarding different items, when, for example, some items concern much debated topics that are lacking information, while other items are simpler and regarding which knowledge is vast and consolidated. It can also be expert-wise $\eta_{ji} = \eta_j$ if it is assumed that experts have different levels of precision in self-evaluating their expertise, or cross expert-item-wise if interaction between the expert and item factors is assumed. The most common distribution for the error is normal with a 0 mean and an $sd\left(w_{j.}\right)/5$ standard deviation. In the case that only a single weight value is used for each expert, the standard deviation of $w_1, \ldots, w_J$ can be used in place of $sd\left(w_{j.}\right)$. The use of 1/5 is in accordance with general practice (see Saltelli & Saisana, 2010). Different values can be used to reflect general higher/lower precision in self-evaluations. A 0 mean reflects neutrality, while in the case that it is assumed that experts are too lenient (strict) in their self-evaluations, a negative (positive) mean is set for the normal error mean (think about, for example, the well-known cognitive bias of overconfidence). Also, distributions other than the normal one can be considered. For example, if very large or small weights are observed with respect to the others, it makes sense to use a distribution with larger than normal tails. A skewed distribution can be used in the case that there is reason to assume the error term distribution is not symmetric.

An alternative way to model uncertainty related to weights is by applying a multiplicative random error:

$${}_2u\left(w_{ji}\right) = w_{ji}\ {}_{ji}, \forall j, i.$$

We suggest generating β_{ji} from a beta distribution defined in $\left[a_{ji}, b_{ji}\right]$, where $0 < a_{ji} < b_{ji}$. For example, by setting $a_{ji} = 0.75$ and $b_{ji} = 1.25$, we can model weight uncertainty by letting weight vary within $\pm$ 25% of the original values. This approach is very flexible, like the additive one, because a_{ji} and b_{ji} can also be set at different values for expert, item or cross item-expert terms to reflect different precision in the self-evaluation of expertise. Moreover, if lenient (strict) self-evaluation is assumed, $b_{ji} < 1$ ($a_{ji} > 1$) can be set. The beta distribution is also very flexible in terms of modelling the shape of weight distribution. In fact, by changing its two shape parameters, it is possible to generate symmetric, skewed, U-shaped, bell-shaped or J-shaped weight distributions.

3.3 Modelling the accuracy of expert opinions

The accuracy of expert opinions is another important source of uncertainty. $X_{j.} = (X_{j1}, ..., X_{jI})$ is the vector of the assessments of expert j, and since they are estimates, it makes sense to account for their accuracy. The rationale is similar to that behind modelling expert weight uncertainty. The most common approach is to assign an additive random error to expert assessments.

${}_1u(X_{ji}) = X_{ji} + \delta_{ji}, \forall j, i$,where δ_{ji} is a random number generated from a standard normal distribution with a 0 mean and an $sd(X_{j.})/5$ standard error. Values other than 5 as the denominator of the δ_{ji} standard error can be used to reflect higher or lower precision in expert assessments. Positive or negative values for the mean of the error term can be used if it is assumed experts tend to overestimate or underestimate item assessments. In theory, it is also possible to assign different random error means to different item assessments if it is assumed or expected that some items are more prone than others to underestimation or overestimation. Again, it is about the effects of cognitive biases, and an example would be the anchoring bias according to which an expert assessment is influenced by a particular reference/previous estimate. Similarly, it is also possible to assign different random error means to different expert assessments if it is assumed or expected that some experts are more prone than others to underestimate or overestimate assessments. Distributions other than the normal one can be used if there is reason to reflect skewness or heavier than normal tails in addressing expert assessment uncertainty.

The framework we are presenting is very flexible, allowing one to also link expert assessment uncertainty to expert weights. Two examples are as follows:

(i) $sd(\delta_{ji}) = f(w_{ji})$, where f is a non-increasing function of w_{ji}. The rationale is that error terms for experts with larger weights have smaller standard errors to reflect higher precision in item assessments due to higher expertise. On the contrary, experts with smaller weights have larger standard errors to reflect lower precision in item assessments due to lower expertise;

(ii) $sd(\delta_{ji}) = g\left(X'_{.i}\right)$, where g is a function of $X'_{.i}$ the vector of expert assessments for item i within a preliminary round of the Delphi study. The rationale is to assign larger variable error terms (lower precision) to items whose assessment shows low convergence during the Delphi study and smaller variable error terms (higher precision) to items whose assessment shows high convergence. Within this logic, we associate the standard errors with the consensus degree of Delphi.

Similarly to the uncertainty related to weights, a multiplicative approach can be considered to model expert assessment uncertainty. In this case, in each step of the uncertainty analysis we compute.

${}_2u(X_{ji}) = X_{ji}\gamma_{ji}, \forall j, i.$

We suggest generating the multiplied error γ_{ji} from the beta distribution defined in $[c_{ji}, d_{ji}]$, where $0 < c_{ji} < d_{ji}$. If it is assumed there is no systematic under- or overestimation in the expert assessments of the items, set $c_{ji} < 1$ and $d_{ji} > 1$. For example, by setting $c_{ji} = 0.8$ and $d_{ji} = 1.2$, we model assessment uncertainty by letting the assessments vary within $\pm$ 20% of the original values. If underestimation is assumed, set $c_{ji} > 1$. For example, by setting $c_{ji} = 1.1$ and $d_{ji} = 1.3$, we model assessment uncertainty by letting the assessments vary within + 10% and + 30% of the original values. Whereas if overestimation is assumed, setting $d_{ji} < 1$, for example, by setting $c_{ji} = 0.75$ and $d_{ji} = 0.9$, we model assessment uncertainty by letting the assessments vary within -10% and -25% of the original

values. As before, c_{ji} and d_{ji} can be also set differently expert-wise, item-wise or for both the expert and the item, to reflect different assessing accuracy by different experts or different accuracy in assessing different items or an interaction between expert and item factors. As already discussed in Subsection 3.2, the beta distribution is also very flexible in terms of modelling the shape of the expert assessment distribution.

4 An application for the development of Delphi-based scenarios

4.1 The data set

The application of the method presented in Sect. 3 is illustrated by analysing data from a Delphi study on the future of Italian families (see Bolzan, 2018). $I = 41$ items were considered. Each item is a short statement aimed at describing a specific aspect of the future of families. The items are grouped in seven different thematic areas (see Appendix A).

The Delphi study was carried out in three successive rounds through computer-assisted web interviewing (CAWI), using the open source online statistical survey web app LimeSurvey (www.limesurvey.org). Experts were asked to assess items in terms of their evolution and relevance over 10 years (the study started in 2018) using a 0–100 discrete scale in increments of 5, with scores of less than 50 indicating a decrease, scores larger than 50 showing an increase and scores equal to 50 indicating invariance with regard to evolution and relevance over the 10 years. Initially, 32 experts were involved, while $J = 30$ participated in all phases of the study, corresponding to a rather low dropout rate of around 6%. Experts were also asked to self-evaluate their expertise using a similar 0–100 discrete scale in increments of 5 regarding the 7 areas into which the items were grouped. After the last Delphi round, we aimed to combine the opinions of the experts regarding the items in order to rank them from the first to the last in terms of evolution. This ranking allows the grouping of items into a certain number of clusters, which forms the basis for the construction of future scenarios (Di Zio et al., 2021).

4.2 Robust weighted aggregation of expert opinions

In this subsection, we illustrate the practical application of the method presented in Sect. 3 to the data set described in Subsection 4.1. More precisely, we would like to combine expert opinions on item evolution. Among the combining functions listed in Subsection 3.1, we selected the additive function

$$_{(.)1}C_i = \sum_{j=1}^{J} w_j N(X_{ji}) / \sum_{j=1}^{J} w_j$$

and the multiplicative function.

$_{(.)2}C_i = \prod_{j=1}^{J} N(X_{ji})^{W_j}$,while from the normalization functions, we chose the z score.

$N_2(X_{ji}) = \frac{X_{ji} - mean(X_{j.})}{sd(X_{j.})}$,the rank transformation

$$N_4(X_{ji}) = \sum_{h=1}^{I} (X_{ji} X_{jh})$$

and the linear scaling.

$N_6(X_{ji}) = \frac{X_{ji}}{max(X_{j.})}$.

These selections are both objective and subjective. For example, z score N_1 is preferred to its robust version N_3 because the item scores are clustered in rather few different values, leading to not very informative values for the median and the median absolute deviation from the median. We could also have picked N_7 in place of N_6, but we preferred N_6 because it was simpler.

Algorithm 1 was run for the first application based on the following steps:

Algorithm 1

Step 1. Set $m = 1$.

Step 2. Set $i = 1$.

Step 3. Randomly select a number from the set {2,4,6}. Let s be this number.

Step 4. If $s = 2$, set $t = 1$; otherwise, let t be randomly selected from the set {1,2}.

Step 5.1. Set $j = 1$.

Step 5.2. Generate a random number η_j from a normal distribution with a 0 mean and an $sd(w_{j.})/5$ standard deviation.

Step 5.3. Compute ${}_1u(w_{ji}) = w_{ji} + \eta_j$.

Step 5.4. Generate a random number δ from a normal distribution with a 0 mean and an $sd(X)/5$ standard deviation, where X denotes the pooled data set.

Step 5.5. Compute ${}_1u(X_{ji}) = X_{ji} + \delta$.

Step 5.6. Repeat Steps 5.2 to 5.5 for $j = 2, \ldots, J$.

Step 6. Compute ${}_{m(s)t}C_i\left({}_1u(w_{.i}), N_s\left({}_1u(X_{.i})\right)\right) = q_{mi}$.

Step 7. Repeat Steps 2 to 6 for $i = 2, \ldots, I$.

Step 8. Repeat Steps 1 to 7 for $m = 2, \ldots, M$.

Step 9. Compute the ranks of matrix $Q = [q_{mi}]$ row-wise. Denote the resulting matrix by $R = [r_{mi}]$.

Step 10. Compute the 5th, 50th and 95th percentiles of matrix R column-wise.

4.3 Results

The results are displayed in Fig. 1. Dots represent the median ranks; vertical segments represent the 5th–95th percentile uncertainty intervals of the ranks. Items are ranked from the one with the highest median rank in terms of evolution (item 31) to the one with the lowest median rank regarding evolution (item 7).

From Fig. 1, it can be observed that the items located in the extreme positions are characterized by lower uncertainty, while the items that occupy the central positions are characterized by greater (although limited) uncertainty. The reasons may lie in the complexity of the topics that the items refer to, without excluding other sources, such as the way in which the concepts have been expressed in the questionnaire, which could result in difficulties in terms of understanding.

The items showing greater uncertainty—in the central part of the graph—refer to the following areas of the questionnaire: parents, housing and policy and services (see Appendix A). These fields can be identified as being associated with the "public sphere" of the family.

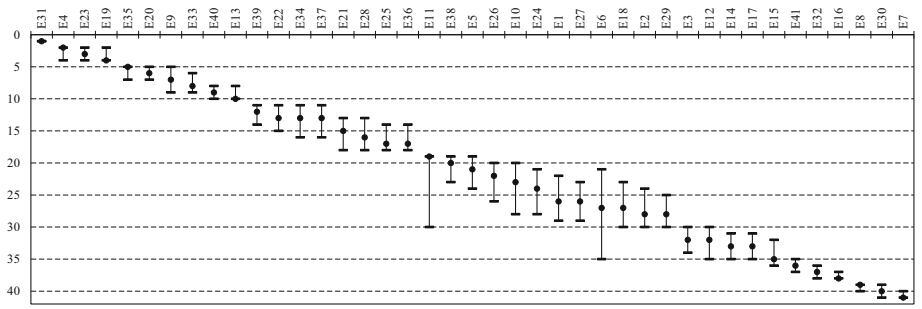

Fig. 1 Robust ranking of items based on their evolution

Conversely, greater robustness of the rank estimates is observed for the items of the following areas: spouses, extended family, children, family models, communication and solidarity. These are issues regarding which the opinions of experts are expressed more immediately, and they recall the idea of the "private sphere" of the family.

The two items showing the least robustness are item 6 (parents will invest in their role as educators of their children) and item 11 (young people will tend to remain in their families of origin once they find employment). Already in the previous phases of the research (see Di Zio et al., 2021), the indicators of the Delphi process had revealed how the experts had experienced some difficulties in terms of reaching a consensus on these two items. One possible explanation is given by the fact that these are issues—particularly in Italy—regarding which family ties are very strong, even in young people who have completed periods of study far from their families of origin (Ambrosini & Rosina, 2009; Bolzan, 2018). In addition, the issue of raising children in Italy still has an inhomogeneous vision due to the strong disparity in actions and interests between the two parents. Therefore, these are two highly debated issues for which it is difficult to reach a consensus, and consequently, the possible future developments are very uncertain.

The previous algorithm was re-run by changing the standard deviation of δ in Step 5.4 to show the effect of different levels of precision in the expert assessments of the items. More precisely, we set $sd(X)/2$ in case (ii) and $sd(X)$ in case (iii). Therefore, we model the expert assessment accuracy with lower precision with respect to case (i), where $sd(X)/5$ was used. Table 1 reports the results for cases (i) to (iii), along with those of case (iv), which will be discussed later.

The results are very interesting. There are very clear patterns in the uncertainty interval extremes as well as in the uncertainty interval lengths item-wise. In particular, we emphasize that.

- The median ranks are mostly the same across cases (i) to (iii) (and (iv), as discussed later);
- The 5th percentiles tend to decrease as the precision in expert assessments decreases;
- The 95th percentiles tend to increase as the precision in expert assessments decreases; and
- The uncertainty interval lengths increase as the precision decreases.

The results for cases (i) to (iii) show that the rankings of items according to expert opinions are most unaffected by assessment precision, whereas rank uncertainty is markedly affected, with higher precision leading to lower uncertainty. The results also show that the method is very useful for what-if analyses. In fact, the researcher can easily change the various settings of the previous algorithm to find out the impact on the results, as we have done here to show the effect of different levels of precision in expert assessments.

Table 1 5th, 50th and 95th percentiles of item rank uncertainty distributions for evolution

Item	Case (i)			Case (ii)			Case (iii)			Case (iv)		
	5th	50th	95th	5th	50th	95th	5th	50th	95th	5th	50th	95th
E1	18	26	33	20	26	30	22	26	29	19	26	31
E2	18	27	33	21	27	31	24	28	30	21	27	31
E3	25	32	37	29	32	35	30	32	34	28	32	36
E4	2	3	4	2	2	4	2	2	4	2	2	4
E5	15	22	30	18	22	28	19	21	24	18	22	27
E6	18	27	37	20	27	36	21	27	35	19	27	37
E7	38	41	41	40	41	41	40	41	41	39	41	41
E8	36	39	41	38	39	41	39	39	40	38	39	41
E9	5	7	11	5	7	10	5	7	9	5	7	10
E10	16	23	31	19	23	29	20	23	28	19	23	28
E11	14	21	30	17	21	29	19	19	30	15	21	30
E12	24	32	37	29	32	35	30	32	35	27	32	36
E13	6	10	15	8	10	11	8	10	10	8	10	11
E14	26	33	37	30	33	36	31	33	35	29	33	36
E15	28	34	38	31	34	37	32	35	36	31	34	37
E16	32	37	40	35	38	39	37	38	38	34	37	39
E17	26	33	38	30	33	36	31	33	35	29	33	36
E18	18	26	33	21	27	30	23	27	30	20	26	32
E19	2	3	5	2	4	4	2	4	4	2	4	5
E20	4	6	10	5	6	8	5	6	7	5	6	8
E21	10	15	23	11	15	19	13	15	18	12	15	18
E22	9	14	21	11	13	17	11	13	15	11	13	17

Table 1 (continued)

Item	Case (i)			Case (ii)			Case (iii)			Case (iv)		
	5th	50th	95th	5th	50th	95th	5th	50th	95th	5th	50th	95th
E23	2	3	5	2	3	4	2	3	4	2	3	4
E24	16	24	32	19	24	29	21	24	28	19	24	30
E25	11	17	25	12	17	20	14	17	18	12	17	21
E26	16	23	31	19	23	29	20	22	26	19	23	29
E27	18	26	33	21	26	30	23	26	29	20	26	31
E28	10	16	23	12	16	19	13	16	18	12	16	19
E29	19	27	34	22	28	31	25	28	30	21	27	31
E30	36	40	41	39	40	41	39	40	41	38	40	41
E31	1	1	1	1	1	1	1	1	1	1	1	1
E32	32	37	40	35	37	38	36	37	38	34	37	38
E33	5	8	12	6	8	10	6	8	9	6	8	10
E34	9	14	21	11	13	17	11	13	16	11	13	17
E35	4	6	10	5	6	8	5	5	7	4	5	7
E36	11	16	24	12	16	19	14	17	18	13	17	19
E37	9	13	21	11	13	17	11	13	16	10	13	18
E38	15	22	30	18	21	27	19	20	23	18	21	27
E39	9	13	20	10	12	16	11	12	14	10	12	16
E40	6	9	14	7	9	11	8	9	10	7	9	11
E41	29	35	39	33	36	38	35	36	37	32	36	38

We will now describe a fourth application, denoted by case (iv), of the method. We aim to showcase the flexibility of the method. Algorithm 2 was run for case (iv) based on the following steps:

Algorithm 2

Steps 1 to 5.3. The same as for Algorithm 1.

Step 5.4. Generate a random number δ_{ji} from a normal distribution with a 0 mean and an $sd(X_{j.})/v(X_{ji})$ standard deviation, where $v(X_{ji}) = \begin{cases} 5 \; if \; X_{ji} \geq 80 \\ 2 \; if \; 60 \leq X_{ji} \leq 75. \\ 1 \; if \; X_{ji} < 60 \end{cases}$

Step 5.5. Compute ${}_1u(X_{ji}) = X_{ji} + \delta_{ji}$.

Steps 5.6 to 10. The same as for Algorithm 1.

With respect to cases (i) to (iii), in case (iv), the precision of the expert assessments is modelled in a much more general manner. The standard error for the item scoring is constant in cases (i) to (iii), whereas both the numerator and denominator of the standard error for the item scoring are not constant. More precisely, the numerator changes as the expert changes; the denominator depends on one's self-evaluated expertise in terms of the classes of items (higher self-evaluated expertise, higher assessment precision, and vice versa) and therefore on changes based on the interaction between the experts and items. In addition, Table 1 shows that for case (iv), the median ranks are mostly the same as before. The level of uncertainty, as indicated by the 5th–95th percentile interval length, is intermediate compared to those of cases (ii) and (iii).

Many other additional analyses were performed. The weight of the experts was also modelled using the multiplicative approach. Many beta distributions with very different shapes were considered along with many different settings for density support. These additional results are not reported because they are very consistent with those corresponding to the additive modelling of experts' weights.

We emphasize, on the one hand, the subjectivity of algorithm setting options, that is the flexibility of the method, which allows the incorporation of the preferences of the researcher (regarding errors, weights …); on the other hand, we have the robustness of the results (median ranks are almost unaffected by the algorithm settings) in addition to the information on the uncertainty, which changes according to the choices made by the researcher. This uncertainty can then become an important input for further analyses and, for example, for constructing future scenarios on families; hence, this is proof that even though the median ranks remain constant, it is necessary to address result uncertainty.

4.4 Future scenario development

As mentioned in Subsection 4.1, the experts evaluated 41 items (i.e. 41 Delphi future projections) in terms of their relevance and evolution. The approach used for the development of the future scenarios is that of Delphi-based scenarios (Di Zio et al., 2021; von der Gracht & Darkow, 2010), where Delphi outputs are grouped in clusters representing draft scenarios. According to our proposal, before grouping the items, it is necessary to create rankings, built as described above, to consider all the sources of uncertainty that occur in a Delphi elicitation.

In many classical methods, the Delphi results are aggregated by choosing one formula for normalization and one for aggregation, but in this way, the resulting future scenarios are

heavily dependent on these subjective choices. In our proposal, on the other hand, the results are robust, precisely because from a Monte Carlo simulation perspective, different formulas are considered simultaneously.

The aggregated judgments, in the form of rankings on evolution and relevance, were used as input in a fuzzy c-means clustering algorithm. This is a non-hierarchical clustering technique (Josien & Liao, 2000) that enables the detection of potential overlaps of items across scenarios. The resulting clusters were then refined by a panel of experts to assess their plausibility and consistency (Di Zio et al., 2021).

For an improved definition of the scenarios in future developments of this approach, it might be useful to consider some measures of dependence among items to ensure the selection of a subset of objects with which most of the information is associated.

5 Conclusion

The main contribution of the method proposed in this paper concerns the integration of both the mathematical and behavioural approaches, exploiting some advantages of both while limiting the disadvantages. By using multiple methods and mixing qualitative and quantitative methodologies, this approach can be included in the mixed-methods category.

It has been shown that uncertainty analysis does not require subjectively selecting a specific mathematical formula to combine expert scores, a normalization formula and a particular weight rule. This is an important point because there is no optimal formula for aggregation and for normalization, nor an optimal weight rule. It has also been shown that uncertainty analysis can model expertise as well as expert scoring accuracy. Another advantage of the proposed method is that its results can be disclosed in a very simple way with an interval graph and are therefore easy to understand, even for stakeholders without mathematical skills.

It has been discussed that a limitation, rather than a disadvantage, of the method is that in its current form, it can be applied to item ranking but not to their ratings. The method is also conceptually applicable with regard to generating ratings, but with far lower flexibility; in fact, ratings are not always comparable when the mathematical formula used varies, unlike rankings, which are always comparable as they are ordinal numbers. An open problem is the full integration of the various qualitative phases within the proposed method. In fact, in this paper, while we also illustrate how it may be possible to integrate the initial phases of Delphi as well, only the results of the last Delphi round are used as inputs for the algorithm. It would be of great interest to explore how both the rankings and the uncertainty intervals change over the course of the different Delphi rounds. Presumably, since the ranks are robust, they may vary slightly, but the variations in the uncertainty intervals throughout the rounds would provide very useful information about the dynamics of consensus development among experts.

The what-if analyses mentioned above could be very useful for the construction of future scenarios. In fact, by modulating the various settings of the algorithm(s) differently, the distinct results in terms of uncertainty intervals provide valuable material for the experts and for the research team in the subsequent steps of evaluation and the refinement of the future scenarios.

Finally, we would like to point out that in future developments of the proposed method it would be useful to take into account the assignment of different weights not only to the experts but also to the items, according to the possible different objectives of the study in question.

Funding Open access funding provided by Università degli Studi G. D'Annunzio Chieti Pescara within the CRUI-CARE Agreement.

Appendix A

List of the items used in the Delphi study

Area 1. Parents (six items)

1. Parents (fathers and mothers) will devote themselves to training their children.
2. Fathers will be present during the training and leisure activities of their children (school, sports, associations, etc.).
3. Mothers will be able to organize work and family life to be more present during their children's educational and free time activities.
4. For mothers, the organization of family life will be conditioned by professional rhythms and commitments.
5. Fathers will try to organize their professional commitments according to the organization of family life.
6. Parents will invest in their role as educators of their children.

Area 2. Spouses, extended family, children (seven items)

7. Couples' relationships will be more intense and solid than their current ones.
8. Spouses will try to organize more or fewer moments together throughout the day.
9. Relationships with the extended family (grandparents, relatives, etc.) will be sought and cultivated.
10. Children will develop peer relationships (schoolmates, friends, etc.).
11. Young people will tend to remain in their families of origin once they find employment.
12. The fathers of children born during a second or subsequent union will organize their time to allow them to also take care of the children they have from previous unions.
13. The mothers of children born during a second or subsequent union will organize their time to allow them to also take care of the children they have from previous unions.

Area 3. Housing (five items)

14. Houses will only have the function of housing family members.
15. Houses will meet the needs of hospitality to encourage moments of sharing among family members.
16. Houses will be the preferred location for meetings between family members (extended family) and between friends.
17. The need for family assistance (regarding children and the elderly) will be met by initiatives of collaboration between families in the same neighbourhoods/condominiums.
18. In the planning of housing construction, people will look for solutions that facilitate meetings between families within condominiums, e.g. habitable common areas.

Area 4. Family models (five items)

19. Unmarried cohabiting couples.
20. Mixed couples, i.e. of different nationalities.
21. Couples comprised of persons of the same sex.
22. Families composed of people who share the same home, e.g. the elderly.
23. Reconstructed families, in which at least one of the members of the couple has one or more children from previous relationships.

Area 5. Policy and services (six items)

24. The local community it belongs to (region, municipality) will consider the family (also through its aggregations and associations) an interlocutor in political decisions.
25. The local community it belongs to (region, municipality) will consider the family (also through its aggregations and associations) a resource to be valued.
26. National policy will consider the family a resource to be exploited.
27. Services will be developed (social, health, etc.) that consider the family a privileged recipient.
28. Urban planning will dedicate spaces for families, e.g. public gardens, small squares.
29. Commercial services (food, household appliances, leisure, tourism, etc.) will consider types of services centred on the family rather than on the individual.

Area 6. Communication (five items)

30. Verbal communication between young people will be more or less frequent.
31. 'Virtual' communication (mobile, social networks, etc.) among young people will be frequent.
32. Verbal communication between parents and children will be frequent.
33. 'Virtual' communication between parents and children will be frequent.
34. Meetings between grandparents and grandchildren will be intense.

Area 7. Solidarity (seven items)

35. Informal networks will be developed between families, e.g. time banks, group purchases.

36. Intergenerational solidarity networks (elderly, adults and young people) will be common.

37. Associations and institutions (parishes, non-profit organizations, etc.) will be able to launch initiatives for the family.

38. The family will be an active subject in voluntary initiatives.

39. Volunteering will focus on family needs.

40. During certain circumstances, such as divorce or the separation of spouses, the family will have access to assistance from competent figures (institutions, counsellors, family mediators) before and during the event.

41. When faced with problems resulting from the loss of the job of one of the spouses, the family will be protected and assisted by institutional bodies (legislation, regulations, etc.) in dealing with problems arising from the event.

References

Ambrosini, E., Rosina, A. (2009). *Non è un Paese per giovani. L'anomalia italiana: una generazione senza voce*. Marsilio, Venezia.

Ayyub, B. M. (2001). *Elicitation of Expert Opinions for Uncertainty and Risks*. CRC Press.

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

Bishop, P., Hines, A., & Collins, T. (2007). The current state of scenario development: An overview of techniques. *Foresight, 9*(1), 5–25.

Bojke, L., Soares, M. O., Claxton, K., Colson, A., Fox, A., Jackson, C., Jankovic, D., Morton, A., Sharples, L. D., & Taylor, A. (2021). Developing a reference protocol for structured expert elicitation in health-care decision-making: A mixed-methods study. *Health Technology Assessment, 25*(37), 1–124.

Bolzan, M. (2018). *Domani in Famiglia: Possibili scenari fra 10 anni*. Franco Angeli Editore, Milano.

Bonaccorsi, A., Apreda, R., & Fantoni, G. (2020). Expert biases in technology foresight. Why they are a problem and how to mitigate them. *Technological Forecasting & Social Change, 151*, 1–17.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.

Colson, R. A., & Cooke, R. M. (2018). Expert elicitation: Using the classical model to validate experts' judgments. *Review of Environmental Economics and Policy, 12*(1), 1–21.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.

Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science, 9*, 458–467.

de Jouvenel, B. (1967). *The Art of Conjecture*. Weidenfeld and Nicolson.

Di Zio, S., Bolzan, M., & Marozzi, M. (2021). Classification of Delphi outputs through robust ranking and fuzzy clustering for Delphi-based scenarios. *Technological Forecasting and Social Change, 173*, 121140.

Di Zio, S., & Pacinelli, A. (2019). Sul controllo della dinamica delle opinioni e della stabilità dei desiderata. In A. Pacinelli, M. Gerarda, & N. Cipolla (Eds.), *Scenari e Partecipazione* (pp. 32–49). Franco Angeli.

Di Zio, S., & Staniscia, B. (2014). Citizen participation and awareness raising in coastal protected areas. A case study from Italy. In A. Montanari (Ed.), *Mitigating Conflicts in Coastal Areas Through Science Dissemination: Fostering Dialogue Between Researchers and Stakeholders* (pp. 155–197, Cap. 6, Vol. 7.), Sapienza Università Editrice, Rome.

Fritschy, C., & Spinler, S. (2019). The impact of autonomous trucks on business models in the automotive and logistics industry – A Delphi-based scenario study. *Technological Forecasting and Social Change, 148*, 119736.

Glenn, J. C., Gordon, T. J. (2009). *Futures Research Methodology* – Version 3.0. The Millennium Project, American Council for the United Nations University, Washington, DC.

Gordon, T. J., Helmer, O. (1964). *Report in a Long-Range Forecasting Study*, RAND corporation Technical Paper P-2982, RAND, Santa Monica, CA, USA.
Gordon, T. J., Sharan, Y., & Florescu, E. (2015). Prospects for Lone Wolf and SIMAD terrorism. *Technological Forecasting and Social Change, 95*, 234–251.
Hanea, A. M., Nane, G. F., Bedford, T., French, S. (2021). *Expert Judgement in Risk and Decision Analysis*. (International Series in Operations Research & Management Science, 293) Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-030-46474-5.
Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2018). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution, 9*(1), 169–180.
Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50*(6), 1141–1151.
Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.
Josien, K., & Liao, T. W. (2000). Integrated use of fuzzy c-means and fuzzy KNN for GT part family and machine cell formation. *International Journal of Production Research, 38*, 3513–3536.
Kosow, H., Gaßner, R. (2008). *Methods of Future and Scenario Analysis: overview, Assessment, and Selection Criteria*. Deutsches Institut für Entwicklungspolitik, Bonn.
Larréché, J. C., & Moinpour, R. (1983). Managerial judgment in marketing: The concept of expertise. *Journal of Marketing Research, 20*(2), 110–121.
Linstone, H. A., & Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. Addison-Wesley Publishing Company.
Marozzi, M. (2021). Perceived justifiability towards morally debatable behaviors across Europe. *Social Indicators Research, 153*, 759–778. https://doi.org/10.1007/s11205-020-02490-w
McCauley, C. (1998). Group dynamics in Janis's theory of groupthink: Backward and forward. *Organizational Behavior and Human Decision Processes, 73*(2), 142–162.
Meyer, M. A., & Booker, J. M. (1991). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Academic Press Limited.
Nowack, M., Endrikat, J., & Guenther, E. (2011). Review of Delphi-based scenario studies: Quality and design considerations. *Technological Forecasting and Social Change, 78*, 1603–1615.
O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician, 73*(sup1), 69–81.
Pesarin, F., & Salmaso, L. (2010). *Permutation Tests for Complex Data*. Wiley.
Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15*(4), 353–375. https://doi.org/10.1016/S0169-2070(99)00018-7
Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society Series A, 168*, 307–323. https://doi.org/10.1111/j.1467-985X.2005.00350.x
Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative–qualitative debate: Implications for mixed-methods research. *Quality and Quantity, 36*(1), 43–53.
Saltelli, A., Saisana, M. (2010). *Uncertainty and Sensitivity Analysis of the 2010 Environmental Performance Index*. EUR 24269 EN. Luxembourg (Luxembourg): Publications.
Schoemaker, P. J. H. (1995). Scenario planning: A tool for strategic thinking. *Sloan Management Review, 36*(2), 25–40.
Sossa, J. W. Z., Halal, W., & Zarta, R. H. (2019). Delphi method: Analysis of rounds, stakeholder and statistical indicators. *Foresight, 5*, 525–544. https://doi.org/10.1108/FS-11-2018-0095
Szwed, P. (2016). *Expert Judgment in Project Management: Narrowing the Theory-Practice Gap*. Project Management Institute Inc.
Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.
Usher, W., & Strachan, N. (2013). An expert elicitation of climate, energy and economic uncertainties. *Energy Policy, 61*, 811–821.
von der Gracht, H. A., & Darkow, I.-L. (2010). Scenarios for the logistics services industry. A Delphi-based analysis for 2025. *International Journal of Production Economics, 127*(1), 46–59.