Risk Analysis, Vol. 00, No. 0, 2021 DOI: 10.1111/risa.13772

Balancing the Elicitation Burden and the Richness of Expert Input When Quantifying Discrete Bayesian Networks

Martine J. Barons,[1,*] Steven Mascaro,[2] and Anca M. Hanea[3]

Structured expert judgment (SEJ) is a method for obtaining estimates of uncertain quantities from groups of experts in a structured way designed to minimize the pervasive cognitive frailties of unstructured approaches. When the number of quantities required is large, the burden on the groups of experts is heavy, and resource constraints may mean that eliciting all the quantities of interest is impossible. Partial elicitations can be complemented with imputation methods for the remaining, unelicited quantities. In the case where the quantities of interest are conditional probability distributions, the natural relationship between the quantities can be exploited to impute missing probabilities. Here we test the Bayesian intelligence interpolation method and its variations for Bayesian network conditional probability tables, called "InterBeta." We compare the various outputs of InterBeta on two cases where conditional probability tables were elicited from groups of experts. We show that interpolated values are in good agreement with experts' values and give guidance on how InterBeta could be used to good effect to reduce expert burden in SEJ exercises.

KEY WORDS: Bayesian network; expert elicitation burden; InterBeta; uncertainty; structured expert judgment

1. INTRODUCTION

Modeling real-life problems often leads to high-dimensional dependence or causal modeling of uncertain variables. Bayesian networks (BNs) are an established type of probabilistic graphical model that provides an elegant way of expressing the joint behavior of a large number of interrelated variables. BNs have been successfully used to represent uncertain knowledge, in a consistent probabilistic manner, in a variety of fields (Weber et al., 2010). They have the advantage that they are transparent with respect to the information used to formulate a response and, when used for decision support or policy evaluation, are also transparent with respect to the decision-making process itself. Being based purely on a probability model, BNs have agreed semantic meanings. A BN is a multivariate statistical model for a set of random variables comprising a directed acyclic graph (DAG) and a set of conditional independence statements. The DAG captures the qualitative structure of the system being modeled, by using vertices/nodes to represent the variables and edges/arcs to indicate statistical dependence between the variables. Each node in the graph corresponds to a random variable and the edges represent direct qualitative dependence/causal relationships. The absence of edges implies a set of (conditional) independence facts. Edges in the BN point from parents (predecessors) to children (successors). A marginal distribution is specified for each node with no parents, and a conditional distribution is

[1]Applied Statistics & Risk Unit, The University of Warwick, Coventry, UK.
[2]Bayesian Intelligence Pty Ltd, Australia.
[3]Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne, Melbourne, VIC, Australia.
*Address correspondence to Martine J. Barons, AS & RU, Department of Statistics, The University of Warwick, CV4 7AL Coventry, UK; Martine.Barons@warwick.ac.uk

0272-4332/21/0100-0001$22.00/1

associated with each child node. These distributions serve as the quantitative information about the strength of the dependencies between the variables involved. A BN can be thought of as a convenient way of representing a factorization of a joint probability mass function or density function of the random variables. The DAG with the conditional independence statements encoded by it, together with the (conditional) distributions, represents the joint distribution over the random variables denoted by the nodes of the graph.

BNs can be discrete or continuous, static, or dynamic. In the case of discrete BNs, the probability distribution at the nodes comes in the form of a conditional probability table (CPT), given the probability for each of the combination of parent nodes on the child node.

Most applications use discrete BNs, i.e., BNs whose nodes represent discrete random variables. Applications involving reasonably rich complexity, often required by practical problems (i.e., child nodes with many parents, discrete variables with many states), require an extremely large number of input values to complete the CPTs. These values can be either retrieved from data, if available, or from experts.

In data-sparse environments, it makes sense to use expert judgment for the quantification. However, an excessive assessment burden on experts may lead to rapid, informal, and indefensible quantification, subject to a raft of cognitive biases. Structured expert judgment (SEJ) elicitation techniques and protocols are available for obtaining estimates of uncertain quantities from groups of experts in a structured way designed to minimize the pervasive cognitive frailties of unstructured approaches. Expert fatigue is a real risk and cognitive loads can be in danger of introducing bias and inconsistency through fatigue. When the quantities of interest are conditional probability distributions, there is a natural relationship between the quantities that can be exploited to reduce the number of items to be elicited, and to complement the elicitation with imputed, analytically calculated inputs. The size of CPTs grows exponentially with the number of parents.

The key question this article aims to address is what is essential to elicit, what can be inferred, and what is the payoff in terms of accuracy. Several methods for eliciting partial CPTs, or completely different measures of dependence that can then be transformed into the needed input (the full CPTs), are available and summarized below.

1.1. Previous Work

In their 2016 systematic review, Werner et al. (2016) raised the question of the burden of elicitation on experts, both in terms of the volume of assessments and also in terms of the complexity of the scenarios. They reviewed methods varying from piecewise interpolation based on the influence of parents (Wisse et al., 2008), to making use of the causal structure in a BN, e.g., the noisy-OR and noisy-MAX methods (Díez, 1993; Pearl, 1988)

Hansson and Sjökvist (2013) compared three elicitation methods with respect to their burden on experts. Taking the battery voltage network in the NETICA software as ground truth, and expert input as required, results from the three methods were compared using the mean absolute difference (MAD) as a measure of accuracy. The three methods used were: a likelihood method, a piecewise linear interpolation elicitation method for Bayesian Belief Nets (called EBBN), and a weighted sum method. For the likelihood method, the expert delivers a typical distribution for the probabilities, a base for the log likelihood, a weighting factor for each state of the child node, and each state of the parent nodes. The EBBN requires the expert to assign as many rows of the CPT as there are child states and one weight for each parent node. The weighted sum algorithm requires the expert to give the relative weights for the parent nodes and the probability distributions for the compatible parent configurations. The likelihood method replicated the originals most closely, as judged by the MAD. It was also the most robust to less smooth probability distributions, but perhaps also represents the highest requirement of technical understanding for the expert.

Alkhairy and Low-Choy (2017) tackle the elicitation burden using alternate methods for selecting the questions to be elicited. The comparison is between Cain's linear interpolation (Cain, 2001), Taguchi's orthogonal arrays (OAs) design (Taguchi, & Konishi, 1987), and a composite of these two. In Cain's method, the rows in the CPT corresponding to "best" and "worst" scenarios[1] are elicited in full and used to construct an integrating factor. One row needs to be elicited for every change in the parent state. This provides a linear interpolation to complete

[1]These often have a very natural meaning in the context of practical problems, for example, the "best" scenario for honey bee abundance is a good environment for food and nesting, average weather, especially normal range temperatures according to season, and low incidence of disease, especially Varroa parasites.

the CPT. Taguchi's OAs design is a fractional factorial method of experimental design in which each pair of parents has the same number of levels and the design is balanced so that all levels from all parents appear in an equal number of scenarios. Rather than a complete factorial design, scenarios are generated by weighted sum generators. The combined method asks experts for the information required by each method, doubling the experts' burden. The Taguchi OA had narrower credible intervals, leading to more accurate estimation of the influences and predictions of quality.

Laitila and Virtanen (2016) refine the methodology for ranked nodes method (RNM). Ranked nodes are used to represent continuous quantities for which there are no well-established ratio or interval scales. Experts provide an aggregation function (weight expression) and weights representing the relative strengths by which the parent node defines the central tendency of the child node on the ordinal scale. The experts also quantify their uncertainty about the parameter by providing a variance. The aggregation function maps points from the normalized scales to the normalized scale of the child node. These points are used as mean parameters of doubly truncated normal distributions, utilized in the generation of the CPT. Laitila and Virtanen (2016) add guidelines for discretizing the interval scales into ordinal scales, guidelines for eliciting the weight expressions, and weights and suggestions for refinement of generated CPT. After discretizing the interval scales of all nodes into an equal number of subintervals, the discretization is refined by asking the expert questions about the subinterval boundary points. Expert assessments about the mode of the child node on the interval scale in various scenarios are used to determine weight expressions and weights. Finally, the expert examines the generated CPTs to ensure that these reflect their views on a series of representative combinations of the states. Nunes et al. (2018) compared the RNM with weighted sum method and a variant of the analytic hierarchy process. For the weighted sum algorithm, experts are expected to deliver compatible parental configurations, i.e., plausible combinations of states. For these compatible parental configurations, experts are asked for probability distributions plus weights for each parent node denoting its degree of influence on the child. The adapted analytic hierarchy process asks the experts for probability assessments conditioned on single parents and calculates the conditional probabilities of nodes with multiple parents. Prior probabilities are obtained by pairwise comparisons of all states of the node, assessing which is more likely and how much more likely. Qualitative descriptions are translated to numerical values. From a reciprocal matrix of these, the relative priority of each state is obtained from the maximum eigenvector. The desired CPTs are calculated using the product of the independent probabilities and a normalizing factor.

In Whitney et al. (2018), similar to the RNM, the authors ask for prior distributions on the parent nodes as a table together with the relative influence of the parents, the effects of each state of the parent node, and the strength of the response. They use the likelihood method described in Hansson and Sjökvist (2013) in their R package "decisionSupport" to calculate CPTs. The parameters were verified with published literature. Then the resulting parameterized BN was shared with experts to verify logical consistency.

The most recent paper is Hassall et al. (2019) in which the authors derive a score for each parent using two questions: the specification of relative parental importance and the associated direction of relationship. Mathematically, this relative weighting and order relationship defines a score, from which an initial draft CPT is created. This score is not designed to fully define a CPT, but rather to provide an initialization that captures the relative effects of the parent nodes while still enabling experts to refine their beliefs through individual edits. The underlying assumptions for this method are that all states can be considered on an equally spaced linear scale and that the range of CPT rows' entries, for a binary child node, will contain values in the full range of 0–100%. If more than two states are necessary for a child node, this scoring approach works best when there is an equidistant definition to the ordinal states of the child node. Although designed purely as an approach to initialize the CPTs, in practice the authors found that relatively little editing of the initialized CPTs was done. Users tended to accept the prepopulated distributions and move on to the next CPT. This was primarily due to either time constraints (with users daunted by the number of tables they needed to complete), or due to low confidence in the elicited relationships (with users opting for a generalized representation of their belief in the absence of any strong feelings to the contrary).

These methods all require different kinds of expertise from the experts involved in the elicitation. The likelihood method requires experts who are

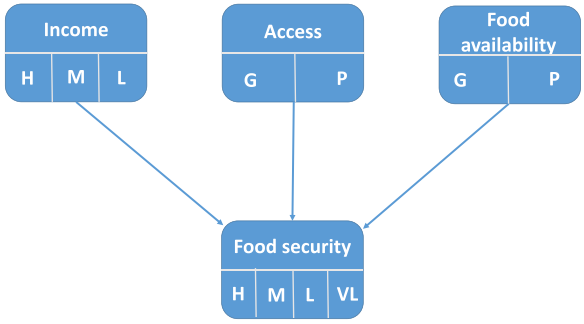

Fig 1. The parents in this BN are physical access, food availability, and equivalized income. Food Security is the child and has four categories. Given all the possible combinations of the parents, the probability distribution CPT requires 48 entries. Since these will become probabilities, we could use the fact that they need to add to one to ask only for 36 probabilities from experts. However, since we were asking for medians, which may not add up to exactly one (plus 90% credible interval), we asked for all 48 entries from each expert and afterward normalized to probabilities.

comfortable with probability distributions and bases for logarithms. The weighted sum method requires probability distributions and also parent weights. The EBBN asks for as many rows of a CPT as there are child states along with parent weights. RNMs require experts to provide an aggregation functions, parent weights, and a variance. The Analytic Hierarchy Process (AHP) requires a probability assessment conditioned on single parents. Hassall et al. (2019) require relative parent importance and the direction of the relationship. After these are used for completing, the CPTs experts are asked to verify and edit the CPTs if necessary. This auditing is also required of experts in the RNM. Cain's linear interpolation requires experts to identify a "best" and "worst" scenario and provide in full the rows relevant to those scenarios plus one row for each change in parent state.

Clearly, then, suitability of these methods will depend very much on the domains of expertise of the experts involved in the elicitation. All experts will be required to have expert knowledge in the domain in which the elicitation exercise seeks to quantify uncertainty. The ability to provide distribution parameters and bases for logarithms requires a good mathematical familiarity in addition to domain knowledge, while the ability to estimate probabilities or natural frequencies—while still challenging for some—has a lower bar.

In this research, we used and tested the Bayesian intelligence interpolation method and its variants for CPTs called "InterBeta" (Mascaro and Woodberry, 2020). We compare the outputs of InterBeta using two case studies where full CPTs were elicited from groups of experts using the IDEA protocol. These experts varied in their mathematical familiarity, so the elicitations were carried out by asking for natural frequencies in one case and probabilities in percentages in the other.

InterBeta focuses on a specific type of relationship between a child and its parents in the BN, namely, ones in which parents influence the parameters of a child's (discretized) Beta distribution. Beta distributions are bounded and can represent unimodal distributions (with a mean and dispersion); thus, this allows modeling cases in which parents affect the mean value of the child, both in linear and nonlinear ways, coupled with noise or uncertainty. This type of relationship is common both when working with continuous variables and also when working with discrete variables that have an underlying continuous nature. The Beta is also flexible enough to support both bimodal and uniform distributions as well, which provides additional flexibility, though we do not examine these possible uses here.

The next section introduces SEJ and the IDEA protocol. We then introduce the technical details relating to the InterBeta methodology, which we will

Table I. Variations on the InterBeta Method Examined Here

	Capability			
Method (Expert Inputs)	Distinct Effects	Nonlinear Effects	Dependencies	Unconstrained Betas
Best and Worst				
Parent Weights	Yes			
Parent State Weights	Yes	Yes		
Row Weights	Yes	Yes	Yes	
Row Beta Parameters	Yes	Yes	Yes	Yes

Note: Every method requires ordered states for parents and child and best and worst case distributions. Some require additional inputs in the form of weights or parameters. Methods lower in the table require more inputs, but allow for more capabilities and therefore greater fidelity.

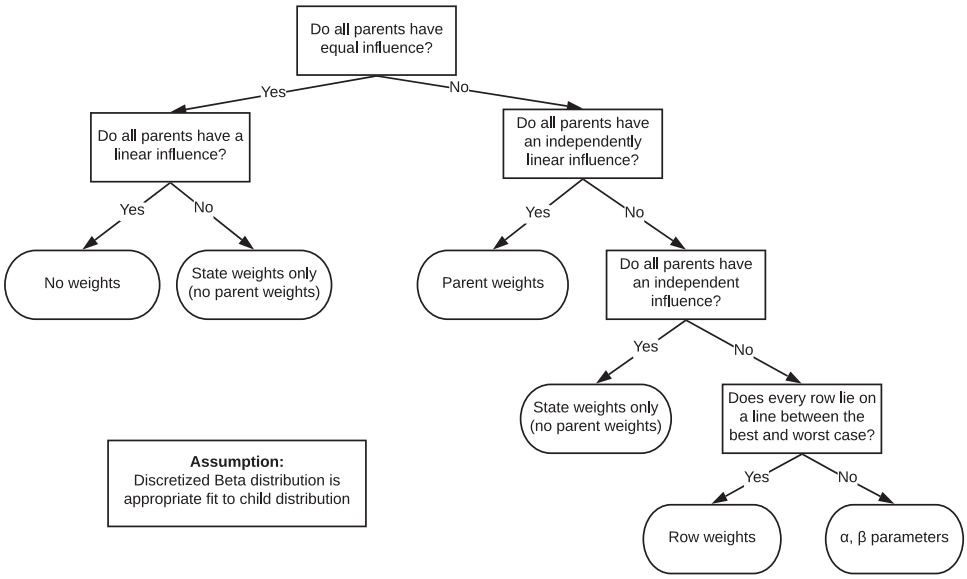

Fig 2. Decision tree for choosing an interpolation method. Note that the burden of each method needs to be balanced against available resources, so an appropriate level of approximation needs to be considered as well. (e.g., In the first question, do all parents have *approximately* equal influence?)

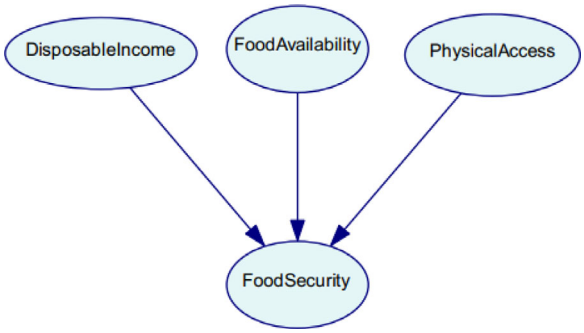

Fig 3. Experts were asked to estimate the food security status (High, Marginal, Low, Very Low) of households given high, moderate or low household disposable income, high, moderate or low food availability, and good or poor physical access.

Table II. Food Example Using the Kullback–Leibler Measure: PW Parent Weights, SW Parent State Weights, RW Row Weights, RP Row Beta Parameters

Expert	PW-KL	SW-KL	RW-KL	RP-KL
Expert A	0.05	0.04	0.02	0.01
Expert B	0.05	0.04	0.02	0.012
Expert C	0.13	0.12	0.10	0.03
Expert D	0.07	0.06	0.04	0.01
Expert E	0.07	0.04	0.03	0.006
Equal Weighted Experts	0.05	0.03	0.02	0.01
Performance Weighted Experts	0.05	0.03	0.02	0.01

Note: Tables showing equivalent Mean Squared Distance (Table A2), Total Variation Distance (Table A3), and Hellinger distance (Table A4) are in the Appendix.

test using the data. Then we give some detail on the measures of performance we use to compare how well InterBeta replicated the original CPTs, given by experts, from partial information. In Section 3, we introduce our data sets and give the results of using the various settings of InterBeta with our data. Finally, we discuss the results in context of reducing expert elicitation burden.

2. METHODS

The InterBeta software was used to complete partial CPTs in BNs using partial information required by the method derived from the full CPTs and imputing the remainder. The imputed values were compared using Kullback–Leibler, mean squared deviation, Hellinger, and total variation distance (TVD) measures, see Section 2.3.

2.1. Structured Expert Judgment

The IDEA protocol (Hanea et al., 2016) for SEJ was used to elicit CPTs from two groups of experts. One elicitation asked for the probability of pollinator abundance, given a range of weather, disease, and environmental conditions. The other SEJ asked for

Table III. Bees Example Using the Kullback Leibler Measure: PW Arithmetic mean, Parent Weights; EW Geometric Mean, Best and Worst; PWG Geometric Mean, Parent Weights; Cain Cain's Method

Expert	PW-KL	EW-KL	PWG-KL	Cain-KL
Expert A	0.028	0.034	0.001	0.002
Expert B	0.016	0.060	0.002	0.002
Expert C	0.047	0.028	0.017	0.023
Expert D	0.002	0.067	0.007	0.004
Expert E	0.02	0.012	0.006	0.008
Expert F	0.001	0.047	0.008	0.007
Expert G	0.023	0.024	0.004	0.005
Expert H	0.016	0.016	0.008	0.009
Expert J	0.002	0.147	0.003	0.003
Expert K	0.032	0.140	0.002	0.001
Equal Weighted Experts	0.006	0.024	0.003	0.002
Average of generated CPTs	0.005	0.027	0.002	0.001

Note: Tables in the Appendix give the equivalent tables for total variation distance (Table A7), mean squared distance (Table A6), and Hellinger distance (Table A8).

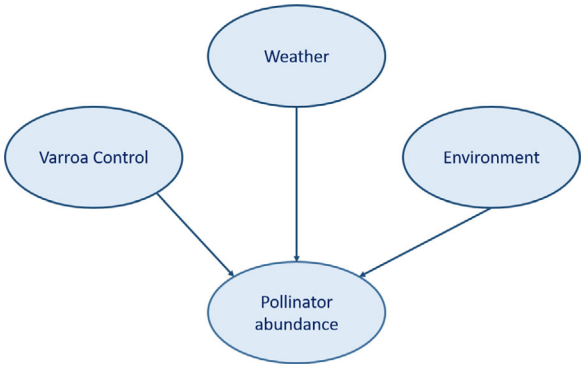

Fig 4. Experts were asked to estimate the abundance of pollinators given good or poor Varroa mite control, average or unusual weather, and supportive or unsupportive environment.

the natural frequencies for membership for four categories of household food security, given a range on household income, access, and food availability conditions (for more details see Section 3.1).

A range of SEJ protocols exist, aiming to subject expert judgment to the same level of care and scrutiny as would be expected for empirical data, ensuring that if judgments are to be used as data, that they are subject to basic scientific principles of review, critical appraisal, and repeatability. Importantly, structured elicitation protocols are grounded in empirical testing to demonstrate that they improve expert judgments.

The IDEA protocol ("Investigate", "Discuss", "Estimate" and "Aggregate"), distills the most valuable steps from existing protocols, and combines them into a single and practical protocol. The protocol has been provided in full detail in Hanea et al. (2016). The key steps are:

- Recruit a diverse group of experts to answer questions with probabilistic or quantitative responses.
- Experts first *Investigate* the questions and clarify their meanings, and then provide their individual best estimates and associated credible intervals in private.
- Experts receive feedback on their estimates in relation to other experts, in anonymized form.
- With the assistance of a facilitator, the experts *Discuss* the results, resolve different interpretations of the questions, cross-examine reasoning and evidence, and then provide a second and final private *Estimate*.
- The individual second-round estimates are then combined using mathematical *Aggregation*.

In our research, we consider eliciting probabilities as relative frequencies. The experts quantify their uncertainty by providing quantiles of the subjective distributions they associate with the elicited relative frequencies. They are typically asked for a best estimate and an upper and lower plausible value interpreted as, for example, 5th and 95th quantiles. If experts' estimates are used to populate CPTs, their best estimates are used as entries in the CPTs and their bounds are used to give an indication of the uncertainty.

2.2. InterBeta Interpolation Techniques

InterBeta (Mascaro and Woodberry, 2020) works with ordered (i.e., ranked) or binary nodes, such as the nodes from Fig. 1. The method requires the user to specify a "best case" distribution and a "worst case" distribution for the child. ("Best" and "worst" here refer to distributions in which the distribution parameters are at their extremes for the child.) All other cases are interpolated between these extremes, based on a set of (noninteracting) weights for the parent states. Note that weights can be assigned to parent *states* not just to parents, hence independently nonlinear (and even nonmonotonic) relationships can be defined. However, interactions are currently limited to weighted sums. (Certain types of interaction can be modeled by introducing

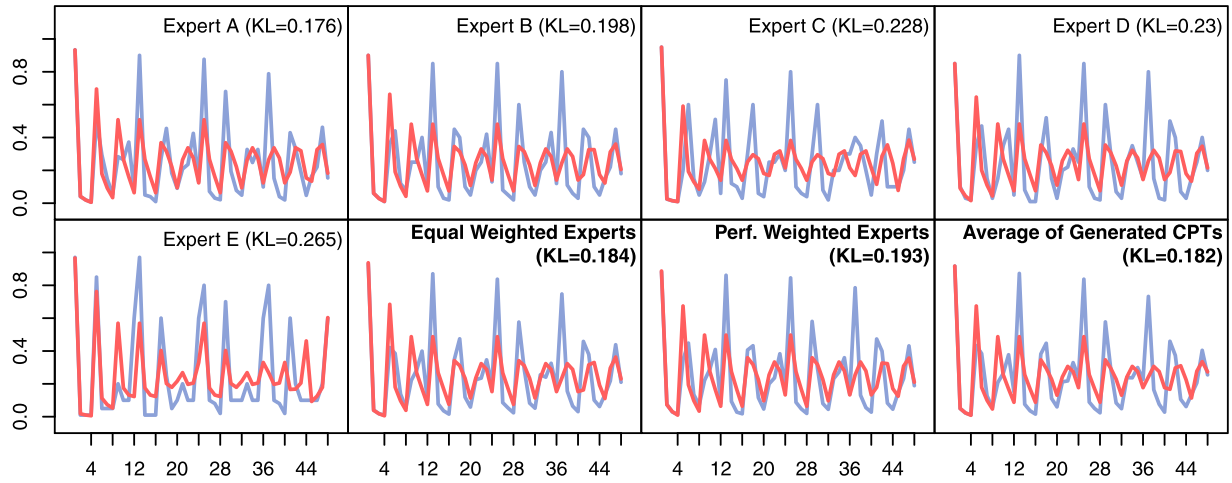

Fig 5. Food Security—Arithmetic Mean—Best and Worst Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line). Each point on the *X*-axis represents a row from the CPT in odometer order (as it would appear in, for example, the Netica BN software package), with the best case at $x = 0$ and the worst case at $x = 47$). (See Figure A4 in the Appendix for ordered graph.) The *Y*-axis is the probability for the question produced by the expert or interpolator. The first five graphs (Experts A–E) represent the CPTs directly given by experts, along with the best fit interpolated CPT. Equal Weighted Experts represents an equal-weighted average of expert CPTs, with the interpolation fit to this average CPT (i.e., the interpolation is run once at the end). Performance Weighted Experts is similar, but weighted instead by expert performance. Finally, Average of Generated CPTs is the same as Equal Weighted Experts, except that the averaging occurs *after* CPTs have been interpolated for each expert (i.e., the interpolation is run for every expert, and then averaged).

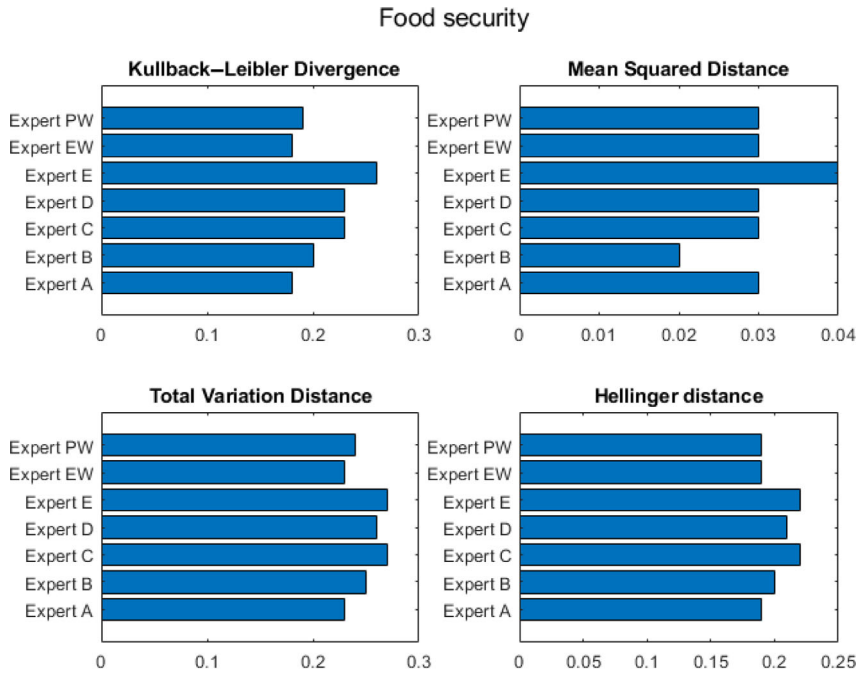

Fig 6. Food security SEJ: Performance measures for Best and Worst interpolation of Food Security SEJ CPTs. Expert EW is the mean of the experts' values and PW Experts is the performance weighted mean of the experts' values. See Table A9.

intermediate nodes, as is already possible in the graphical language of BNs themselves.)

The interpolation techniques we use here assume that the conditional distribution for the child node can be approximated by a Beta distribution. Even though other distributions can be assumed, the Beta distribution is chosen for its bounded support and its flexibility (unimodal, uniform, and bimodal at both extremes). The interpolation works by interpolating the *parameters* of the Beta distribution (rather than

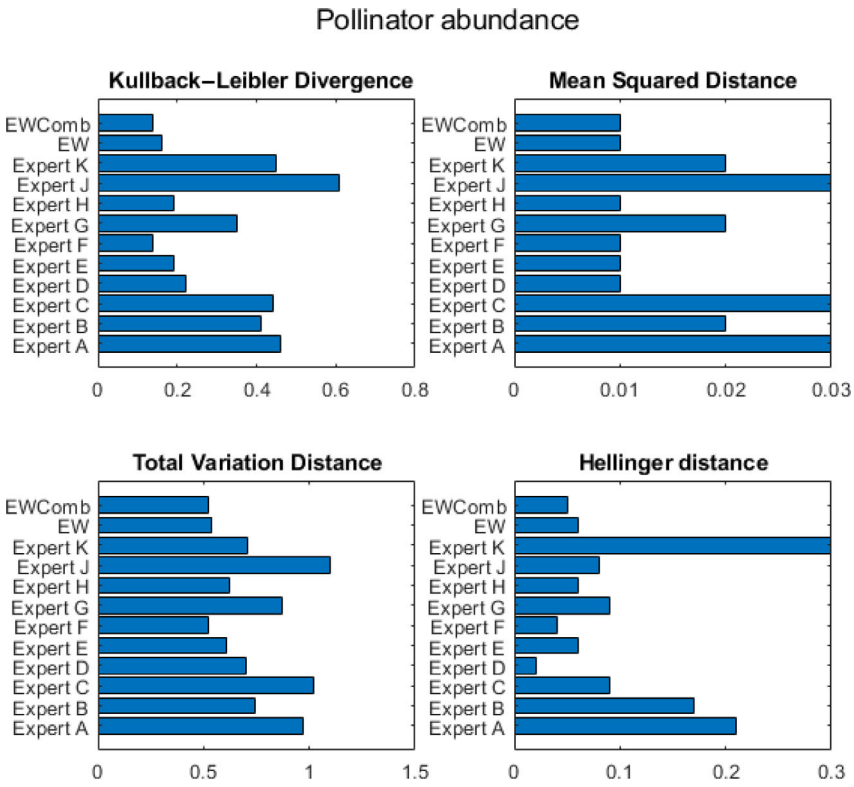

Fig 7. Pollinator abundance SEJ: There was no performance weighting in this data set since the measures of performance were unable to justify performance weighting. Equal weighted comb is the mean of the imputed values for comparison with EW, the mean of the experts' values. See Table A10

the distribution itself), thereby maintaining the shape of the Beta across all the interpolated rows.[2] That is:

$$P(X|y_{1\cdots n}) \sim Beta(g_\alpha(y_{1\cdots n}), g_\beta(y_{1\cdots n})),$$

where X is the child, $y_{1\cdots n}$ is a combination of parent states identifying a single CPT row (for parents $Y_{1\cdots n}$), and g_α and g_β map these parent state combinations to α and β parameter values for the Beta distribution. The user is not required to provide the Beta parameters directly; indeed, they need not even know the tool uses Beta distributions at all. At present, the tool allows the user to provide a multinomial with $n = 1$ (i.e., a categorical distribution) or the α and β parameters directly.[3], [4]

[2]This is in contrast to the common case when combining probabilistic expert judgments, which typically involves combining the distributions themselves, rather than the parameters. This approach can work well for expert judgments, but works poorly for interpolating CPTs where the shape of the distribution washes out or becomes distorted for the middle CPT rows.

[3]Currently, the tool elicits multinomials with the assumption that $n = 1$, though in future, we may allow $n > 1$ to capture (for example) an estimate of confidence via equivalent sample sizes. If multinomials are provided, a Beta distribution is fitted to it using a simple stochastic hill climbing search.

[4]Throughout this article, we will use "Beta parameters" as shorthand to refer to any means of specifying a distribution, whether by multinomials, means and dispersions, quartiles, or actual α, β parameters.

Given the Beta distribution is continuous, it must be discretized. The tool treats the states of the target node as ranging over an interval $[0, |X|]$, where $|X|$ is the number of states in the child node. The user specifies their best and worst case distributions over this interval—either with multinomials or Beta parameters. To recover the multinomial for any particular row in the CPT (whether a best case, worst case, or interpolated row), the program calculates the probability mass within each unit interval (i.e., $[0, 1), [1, 2), \ldots, [|X| - 1, |X|]$) and assigns it to the corresponding state.

To determine the parameters for any particular interpolated row, InterBeta takes a weighted mixture of the best and worst case Beta parameters ($(\alpha_\uparrow, \beta_\uparrow)$ and $(\alpha_\downarrow, \beta_\downarrow)$, respectively) as follows:

$$\alpha = g_\alpha(y_{1\cdots n}) = w(y_{1\cdots n})\alpha_\uparrow + (1 - w(y_{1\cdots n}))\alpha_\downarrow,$$
$$\beta = g_\beta(y_{1\cdots n}) = w(y_{1\cdots n})\beta_\uparrow + (1 - w(y_{1\cdots n}))\beta_\downarrow,$$

where $w(y_{1\cdots n})$ is a weight function that maps parent state combinations (or rows) to weights that fall between 0 and 1. Weights can be supplied for each row

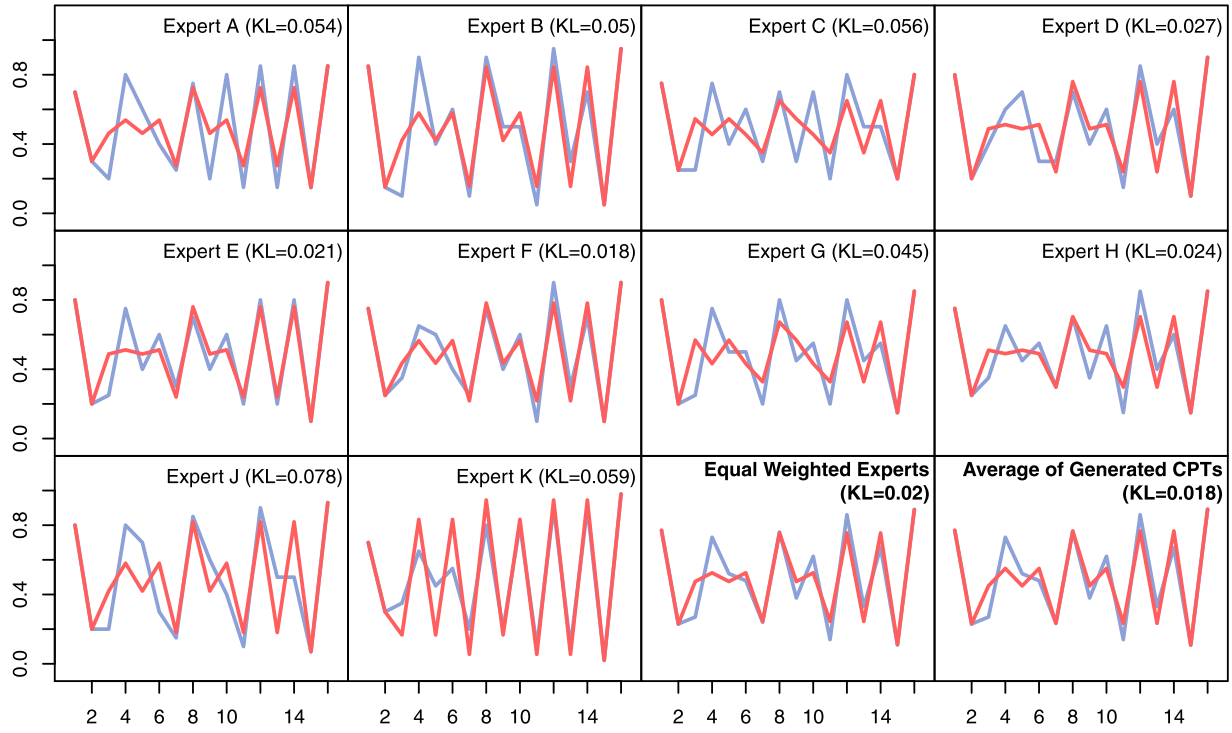

Fig 8. Bees—Arithmetic Mean—Best and Worst Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the x-axis and probability on the y-axis.

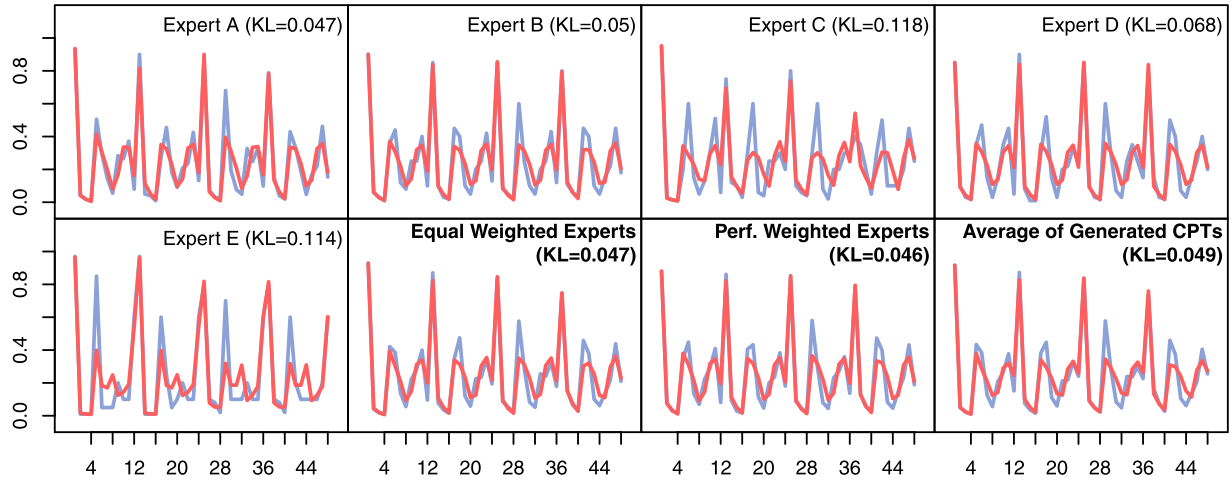

Fig 9. Food Security—Arithmetic Mean—Parent Weights Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the x-axis and probability on the y-axis.

directly, or more commonly computed from the parent states. This can be a simple linear combination of the parent states, or some other combination, such as the arithmetic, geometric, or harmonic mean. In this article, we focus just on the arithmetic and geometric means. As an alternative to interpolating the α and β parameters themselves, we could instead interpolate the mean and variance. Of these two approaches, our experiments have not yet suggested a clear winner with respect to how well CPTs fit their originals, and therefore, we focus here on our original interpolation method that is applied to the α and β. However, it

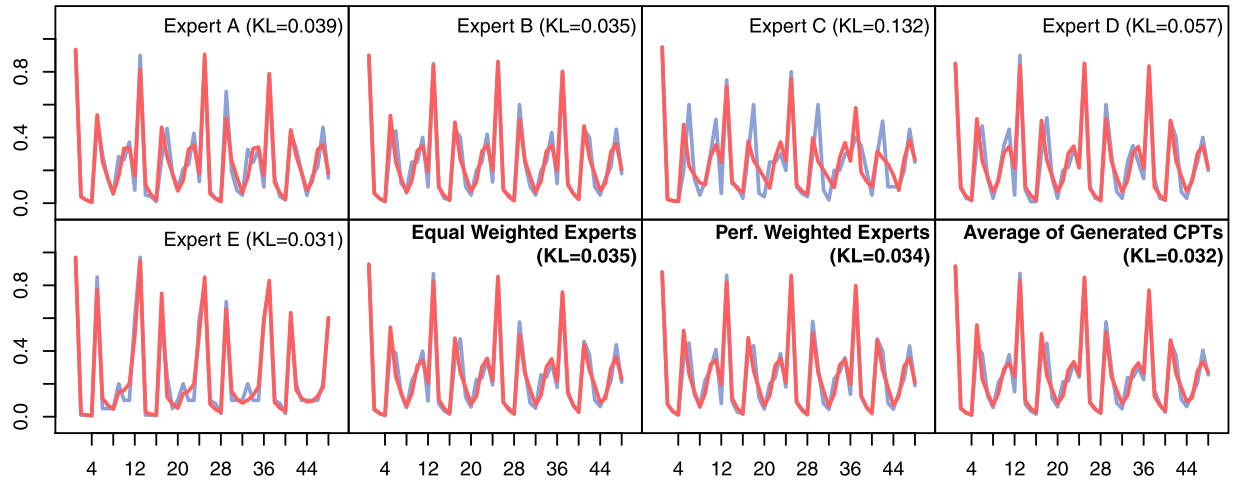

Fig 10. Food Security—Arithmetic Mean—Parent State Weights Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis

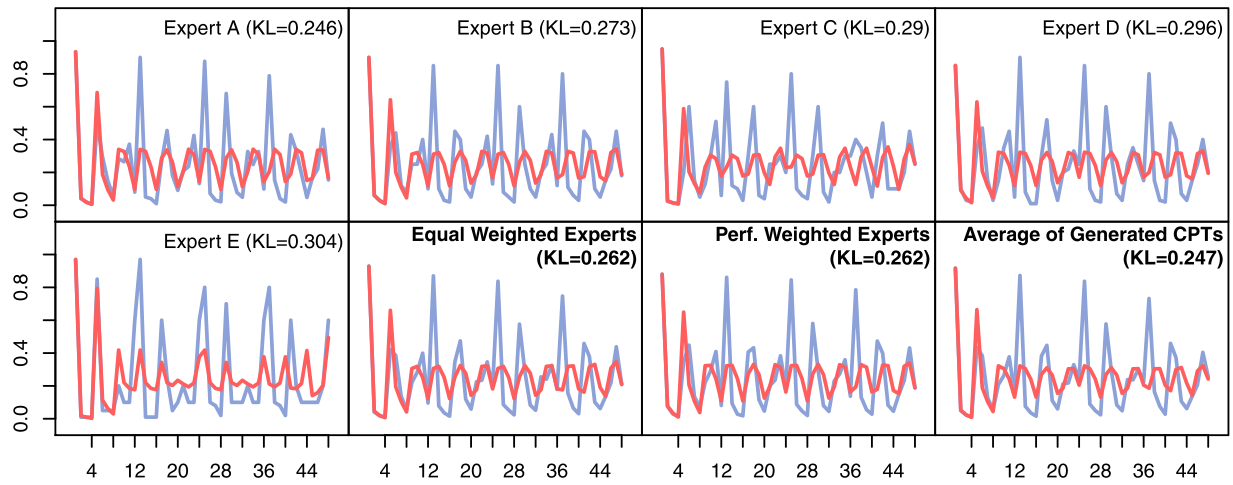

Fig 11. Food Security—Geometric Mean—Best and Worst Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis.

may be that the interpolation of means and variances produces CPTs that better align with expert expectations.

For example, consider the Food Security node in Fig. 1 and the pictured (and empty) CPT. Suppose an expert has provided us with the best case distribution of [0.5,0.2,0.2,0.1], corresponding to the top row (where physical access and food availability are both good, and equivalized income is high). Our stochastic hill-climbing search finds an approximate of $\alpha_{\uparrow} \approx 0.6$ and $\beta_{\uparrow} \approx 1.2$. The expert also provides a worst case distribution of [0.1,0.2,0.3,0.4] (for the bottom row, where physical access and food availability are both poor, and equivalized income is low), and this is fitted with $\alpha_{\downarrow} \approx 1.6$ and $\beta_{\downarrow} \approx 0.9$. The best case row would be given the maximum weight (of 1) and the worst case row the minimum weight (of 0), and these would therefore result in just the degenerate interpolations equaling the best and worst Beta parameters. For an intermediate row, the weight would fall somewhere in between. Say, for example, a row close to the best case has a weight of 0.85. Based on the equations above, the interpolation for this row would give:

$$\alpha = g_{\alpha}(y_{1\cdots n}) = 0.85 \times 0.6 + (1 - 0.85) \times 1.6 = 0.75,$$
$$\beta = g_{\beta}(y_{1\cdots n}) = 0.85 \times 1.2 + (1 - 0.85) \times 0.9 = 1.155,$$

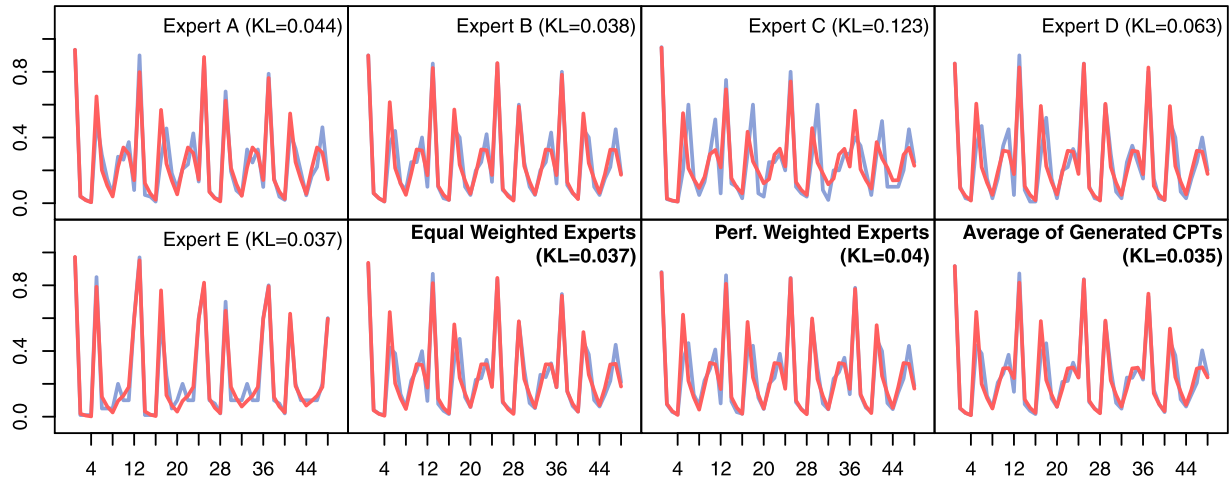

Fig 12. Food Security—Geometric Mean—Parent Weights Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis.

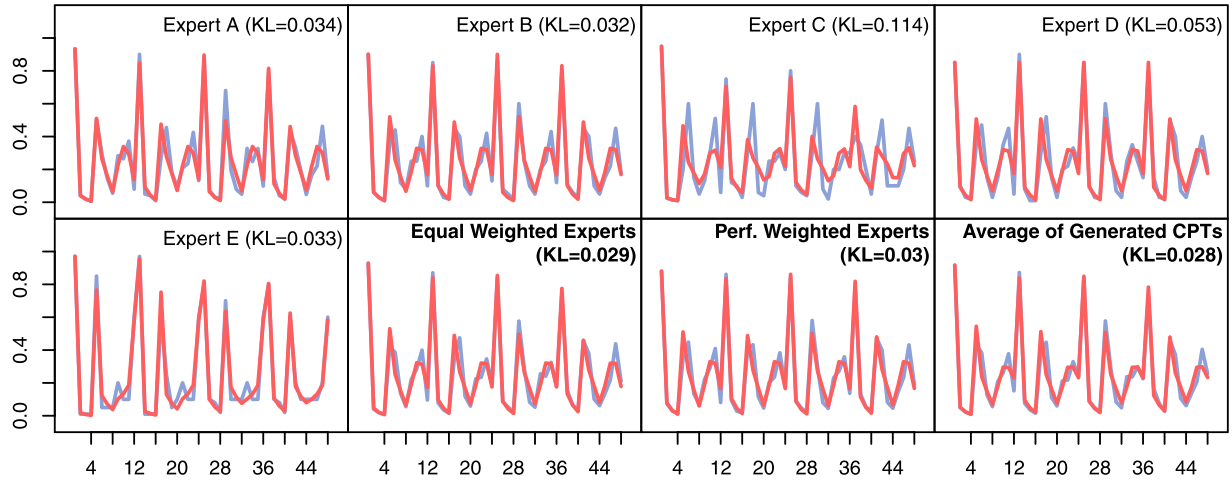

Fig 13. Food Security—Geometric Mean—Parent State Weights Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis.

giving a distribution (prior to discretization) of $Beta(0.75, 1.155)$, which is slightly flatter and skewed more to the right of the best case approximate distribution of $Beta(0.6, 1.2)$. As the row weights move closer to 0, the mass of the distribution shifts further to the right and closer to the worst case approximate distribution of $Beta(1.6, 0.9)$.

While InterBeta works exclusively via Beta distributions at present, there is nothing that prohibits the same interpolation approach being used with other types of distribution (such as truncated normals, log normals, triangular distributions, or even general multinomials). In addition, while we focus here on cases in which the user supplies two CPT rows (the best and worst case), it is also possible to extend the approach to multirow interpolations. (See Mascaro and Woodberry (2020) for further discussion.)

2.2.1. *Variations*

There are several variations on the interpolation method, ordered based on the number of required inputs in Table I. The number of inputs is positively correlated with how well the interpolation can replicate an arbitrary CPT—hence, the more inputs we have, the closer the interpolation will match an expert's fully specified CPT. For all interpolation

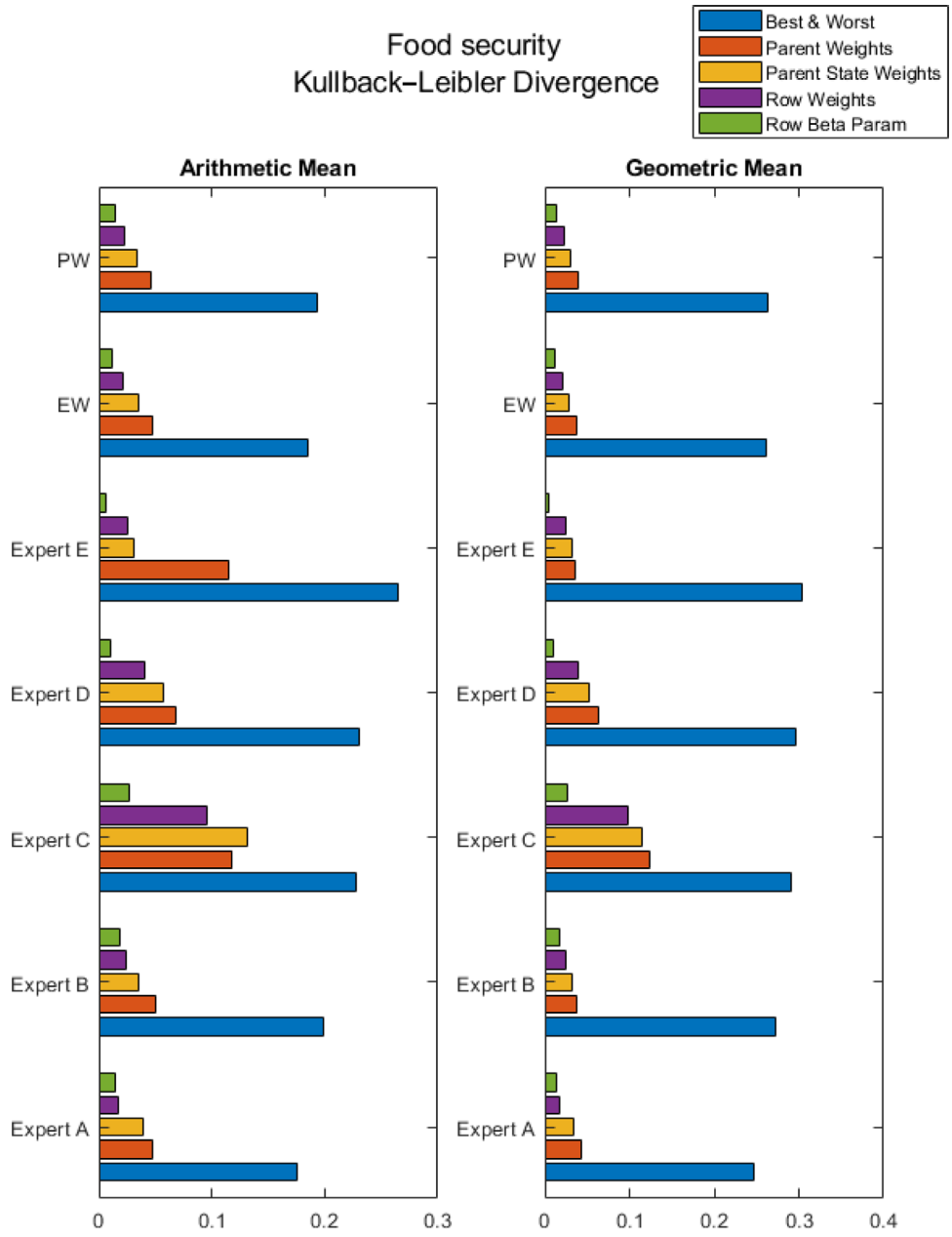

Fig 14. Food Security—Kullback–Leibler—Arithmetic and Geometric Mean—Comparison. For other distance measures, see Figs. A6–A8.

variations specified in the table, the user must at least specify both a best case and worst case Beta distribution for the child, as well as a state order for each node (which may be just the natural state order for each node). In addition, the user *may* be asked for parent weights, weights for all parent states (separately for each parent), or weights for all rows (i.e., all parent state combinations); alternatively, they can specify the Beta parameters for all rows.

If only the best and worst case distributions are supplied by the user (Best and Worst), the parents (y_i) are assigned the same weight (w) for the interpolation. If a parent contains two states, the top state is assigned the full parent weight, while the bottom state is assigned 0 weight. If a parent contains more than two states, the intermediate states receive a partial weight, with uniform spacing between states: for a parent with n states, the top state receives

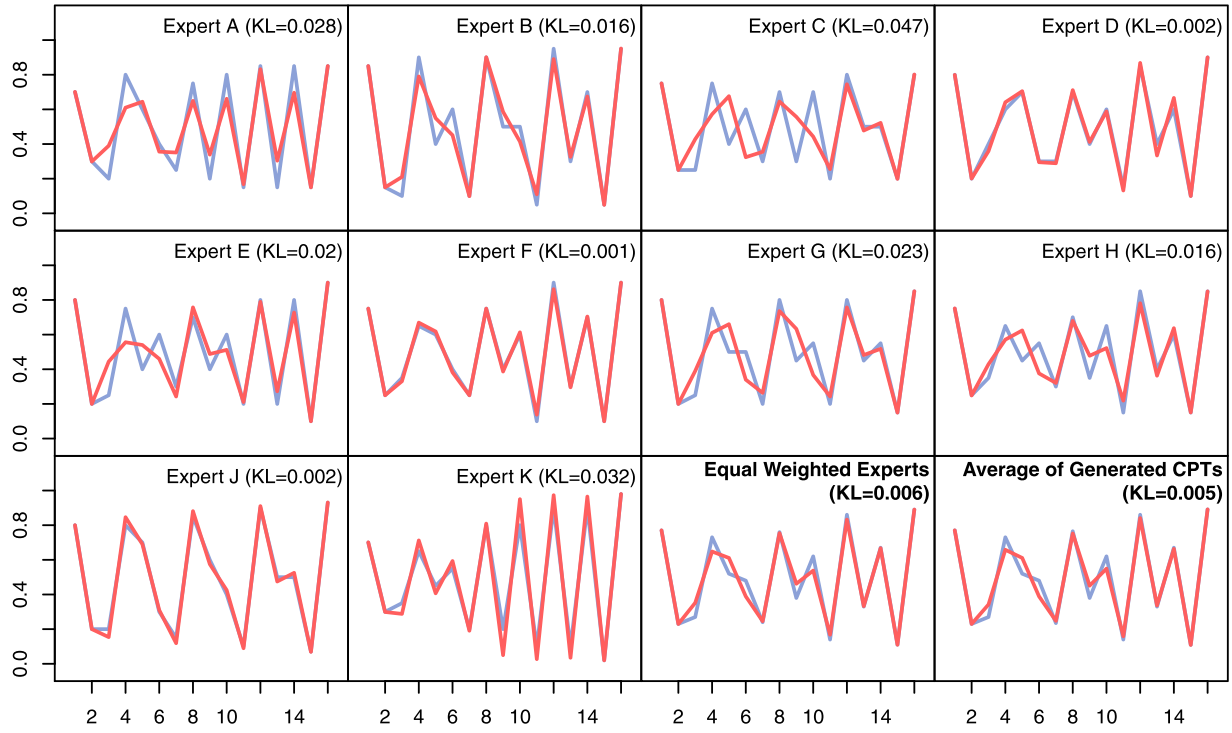

Fig 15. Bees—Arithmetic Mean—Parent Weights Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the x-axis and probability on the y-axis.

weight $\frac{n-1}{n-1}w = w$, the next receives $\frac{n-2}{n-1}w$, the next receives $\frac{n-3}{n-1}w$, and so forth until the last, which receives $\frac{n-n}{n-1}w = 0$.

The user can also specify parent weights w_{Y_i} (Parent Weights), allowing the effects of parents to be distinct (i.e., one parent can have more influence than another).[5] The user can also specify weights for parent *states* $w_{Y_i=y_i}$ (Parent State Weights). Again, there is no change to the dependencies that can be captured in this way, but nonlinear and nonmonotonic relationships can now be captured. Note that whether weights are specified for parents or their states, parents can still only have independent effects on the child.

Taking again the Food Security example described in the previous section, suppose that the expert has suggested that all parents should receive equal weight, and all states are also of equal weight. Arbitrarily, we can assign a weight of 1 to every parent. For physical access and food availability, the full weight of 1 would go to the state "Good," while the state "Poor" would receive weight 0. For equivalized income, the full weight of 1 would go to "High," 0.5 would go to "Moderate" and 0 to "Low." To compute the weight for a given row, we can then use one of our combination methods on the weights associated with each state. For example, if we compute the weight using the arithmetic mean for the row where physical access is poor (state weight of 0), food availability is good (state weight of 1), and equivalized income is moderate (state weight of 0.5), we get $(0 + 1 + 0.5)/3 = 0.5$. This interpolation weight will yield a distribution much closer to the best case than the worst case (as described in the previous section).

Specifying weights (Row Weights) or Beta parameters (Row Beta Parameters) for all rows may seem redundant, given the number of elicited parameters will be equal to or greater than the number of CPT rows; however, this can still be useful for: (1) elicitation, when the child has a large number of states; (2) capturing the expert's intent (particularly for Row Weights, where weights are directly interpretable as the relative distance to the best and worst cases); (3) aggregating multiple expert inputs;

[5]Nondistinct parents can be useful when dealing with homogeneous parents that may vary in number; e.g., providing a summary for the health condition of a set of a trees. An analogy can be made with the statistical concept of identically distributed variables.

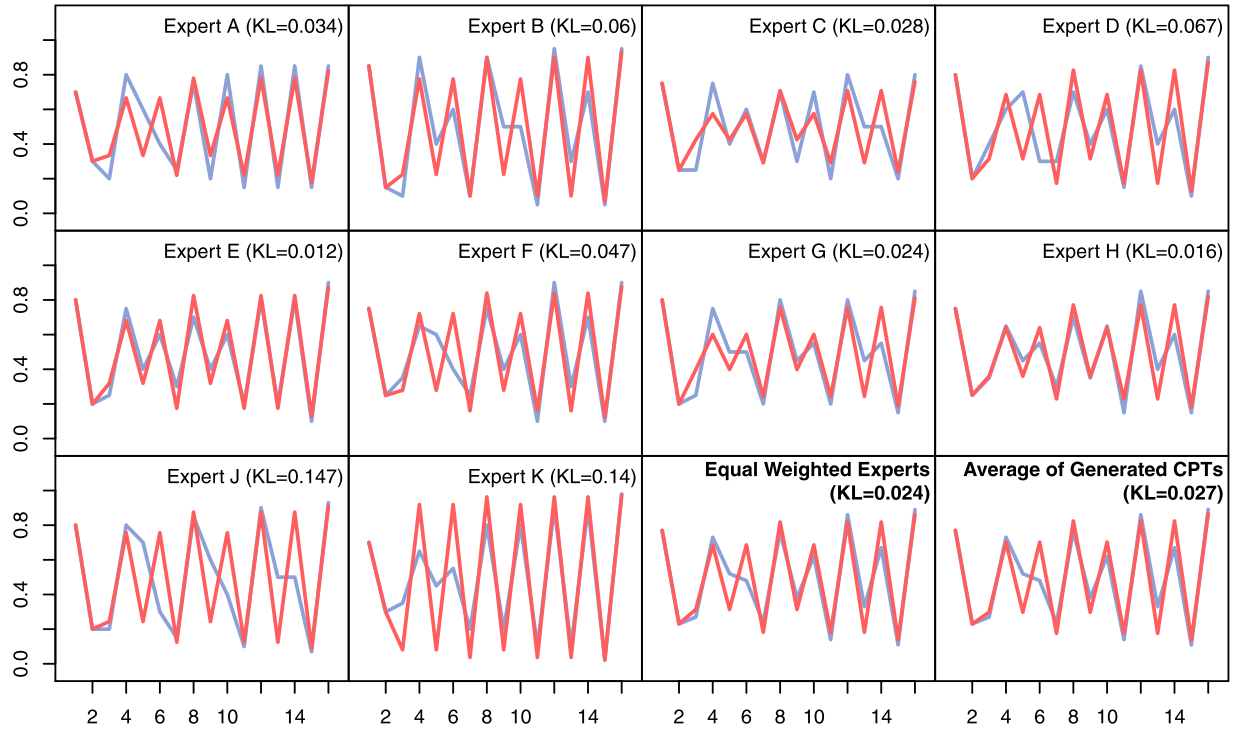

Fig 16. Bees—Geometric Mean - Best and Worst Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis.

or (4) when the child state space needs to be discretized either dynamically or in different ways for different contexts.

Most notably, eliciting row weights allows dependencies between inputs to be captured. For example, suppose that a binary parent Toggle = {on, off} inverts the influence of the other parents. This can be easily represented by assigning row weights in one direction when Toggle is on, and in the reverse direction, when it is off. However, this ability is limited to points that lie on the interpolation line. If, for example, a specific combination of parent states leads to an entirely different Beta distribution, this cannot be captured. If this is required, unconstrained Betas (that is, not constrained to the interpolation line) can be provided for each row instead. While this provides the greatest fidelity out of the above techniques, it also involves the greatest elicitation burden.[6] In most practical cases, experts would need not specify the full set of rows. Much like Hassall et al. (2019), the CPTs can be generated with one of the less burdensome methods first, and then the CPT can be reviewed and modified as needed, but doing the review and adjustment using the generated row weights (or Beta parameters) instead.

[6] As described in Mascaro and Woodberry (2020), specifying row weights only allows direct dependencies between the inputs and the interpolation score node, not directly to the final Beta distribution node.

2.2.2. *InterBeta in an Expert Elicitation Context*

Suppose we want to parameterize a CPT, we have a group of experts that we can elicit from and we have also settled on some structured elicitation protocol.

Prior to using InterBeta, we should confirm that the child variable can be approximated well by a uniformly discretized Beta distribution—that is, will the distribution of the variable always have (approximately) one of the following shapes: (1) uniform (i.e., flat), (2) unimodal, or (3) bimodal, with each maximum located at either end. If the variable satisfies this property, InterBeta can be used.

With this settled, the first step for elicitation would be to decide on the interpolation method. The modeler can use the answers to certain very specific questions to guide this choice; such a guide is provided in Fig. 2. However, the choice of method needs

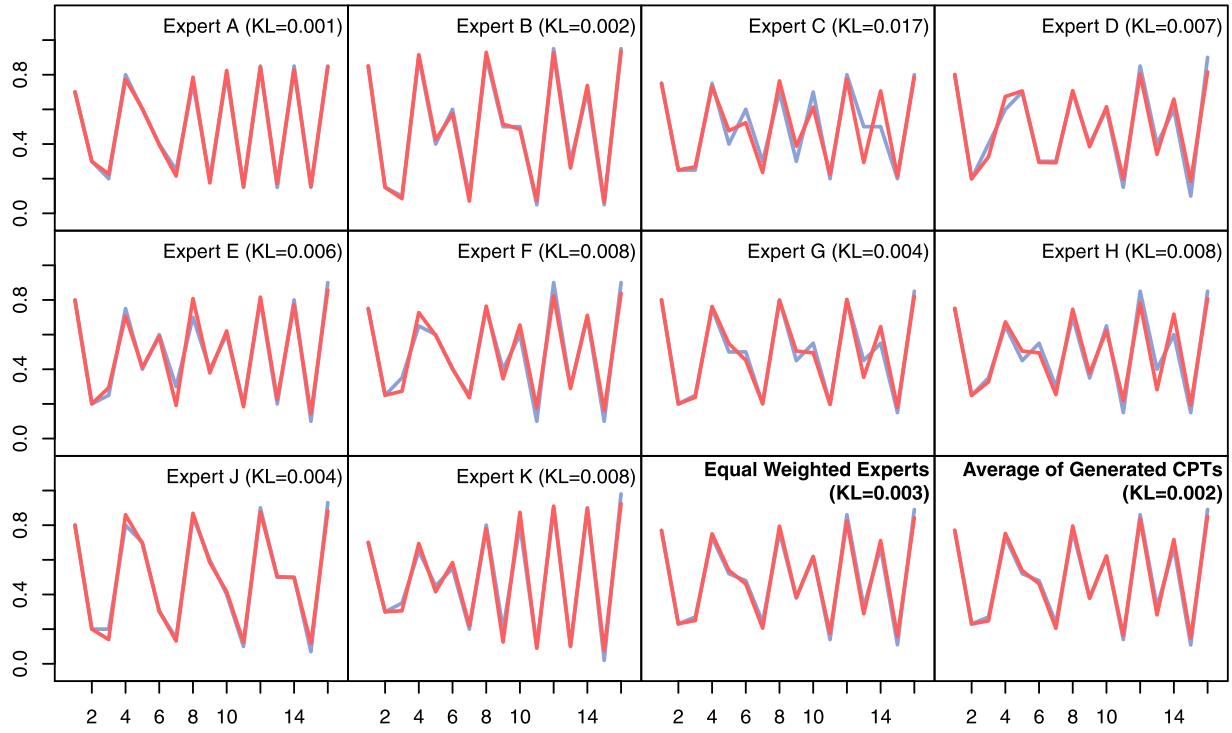

Fig 17. Bees—Geometric Mean—Parent Weights Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis.

to balance other modeling considerations as well—for example, the importance of the node to both the model and its purpose, as well as other resource constraints and problem constraints. This decision might therefore be better taken by the group of experts, assuming that they are familiar enough with what each method entails and consensus can be reached.

If decided by the experts, discussion should be stimulated based on a series of questions of the following sort:

- Do some parents affect the child to a greater extent than others?[7]
- Do some parents affect the child nonlinearly?
- Do certain combinations of parent states affect the child more or less than they would alone (e.g., do the parents affect the child synergistically)?
- Are there important special cases in which the child distribution is very different?
- Does the child (or its parents) significantly influence the rest of the network (and in particular, the key nodes of interest)?

[7]The exact wording for this and other questions would depend on the domain, e.g., Do some of the following demographic and environmental factors affect a person's food security much more than others?

Guided by this discussion and the technical considerations in Fig. 2, the modeler can choose an appropriate interpolation method. For example, if the answer to all of the above questions is a very firm yes, we may need to choose either the row weights or Beta parameters method. Instead, if the answer to the question of synergies is no (or no significant synergies), we have good justification for selecting one of the less burdensome methods, such as Parent Weights or Parent State Weights.

The choice of interpolation method will then guide what needs to be elicited from experts as a basis for interpolation. The method of eliciting those values from experts can be selected from appropriate protocols. Here the IDEA protocol was used which asked for a best estimate (median) and also a lowest and highest plausible value, giving a 90% credible range. The median value was used for the interpolation.

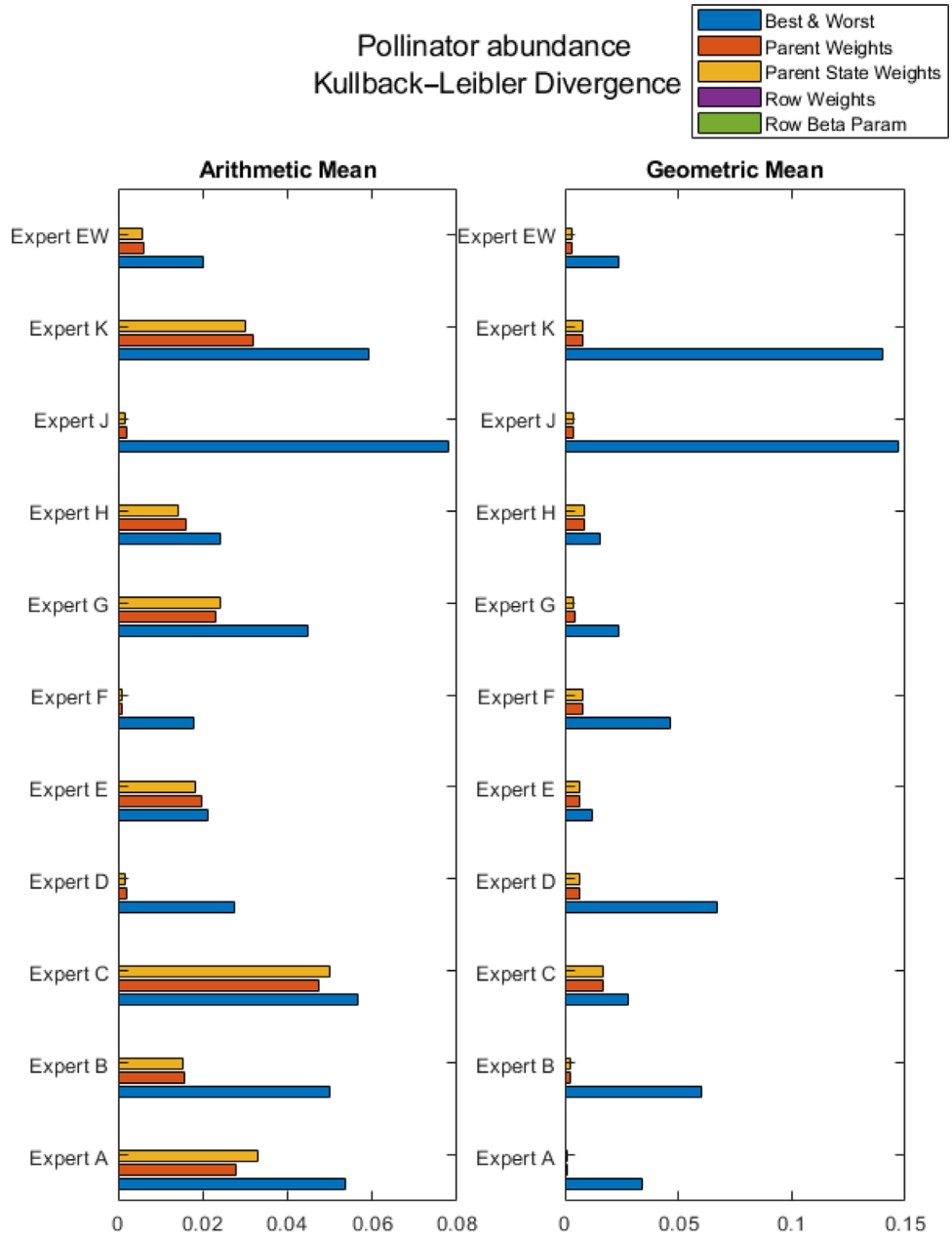

Fig 18. Bees—Kullback–Leibler—Arithmetic and Geometric Mean–Comparison. For other distance measures, see Figs. A9,–A11.

Once the interpolation method is chosen—most typically Parent Weights—we would begin with asking for the best and worst case scenarios (using the selected elicitation protocol), then identifying the most important parent, and then asking how significant the other parents are as a fraction of the most important parent. Below is an illustration, with sample responses in bold:

- Consider the best possible case for all of the given factors. How likely is FoodSecurity to be High, Medium, Low, Very Low?
  Multinomial: [0.9,0.1,0,0]
- Consider the worst possible case for all of the given factors. How likely is FoodSecurity to be High, Medium, Low, Very Low?
  Multinomial: [0.1,0.2,0.4,0.3]

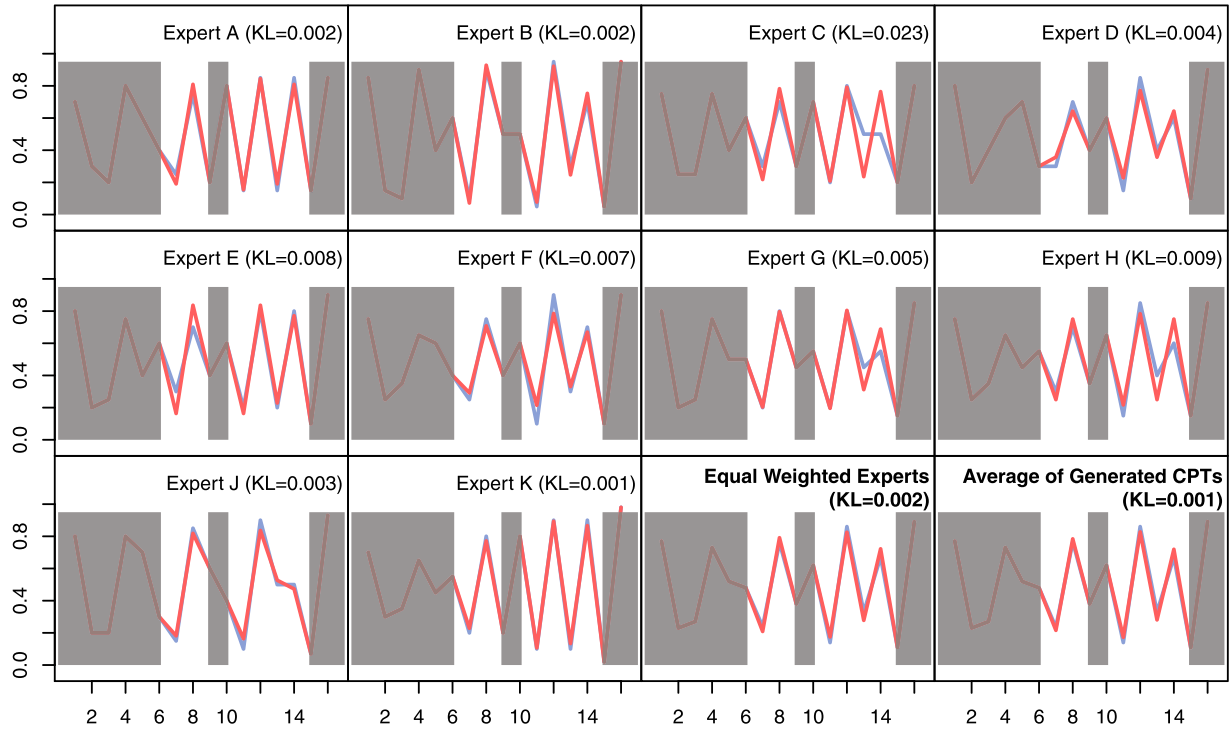

Fig 19. Bees—Cain Comparison of the expert CPTs (blue line) to the interpolated CPTs (red line) with question number on the *x*-axis and probability on the *y*-axis. Parameters 7–8 and 11–14 are the only interpolated parameters, and the rest (shaded out in grey) are given by the expert.

- Which factor is the most important in determining FoodSecurity? (Choose just one.)
  FoodAvailability
- Which factor is the second most important in determining FoodSecurity? To what extent is it less important than the first factor above? (e.g., just as important, 50% as important, 20% as important, etc.)
  PhysicalAccess: 80% as important

The individual distributions need not be elicited as multinomials. Direct parameters can be elicited, such as means and dispersions, or α and β parameters. It should be noted that things like the framing, intelligibility, ordering, etc., of questions affect the responses that experts provide, sometimes significantly. While we do not explore these issues here, we do note that research on elicitation techniques advises against eliciting means and dispersions (let alone more abstract parameters). Best practice in this case would be to ask for 3 quantiles of the distribution (for variables that may have an alternative continuous representation), assuming that the question can be formulated in terms of relative frequency formats.

When weighted rows is the preferred method, there may be additional savings in effort if the row weights are generated by InterBeta first using one of the lower burden techniques (such as Best and Worst) and then adjusted by experts, rather than elicited from scratch. Experts could modify the row weights directly if they are comfortable with their meaning (saving time), or otherwise transformations could be provided (e.g., into Beta distributions, quantiles, or graphs/visual scales) that could instead be the basis for modification, similar to the procedure in the ACE software (Hassall et al., 2019).

Once the expert group's responses have been collected, independent of the elicitation method, the analyst must make a choice about how the interpolation and aggregation interact. There are two main possibilities:

1. Interpolate individual responses, then aggregate.
2. Aggregate responses, then interpolate.

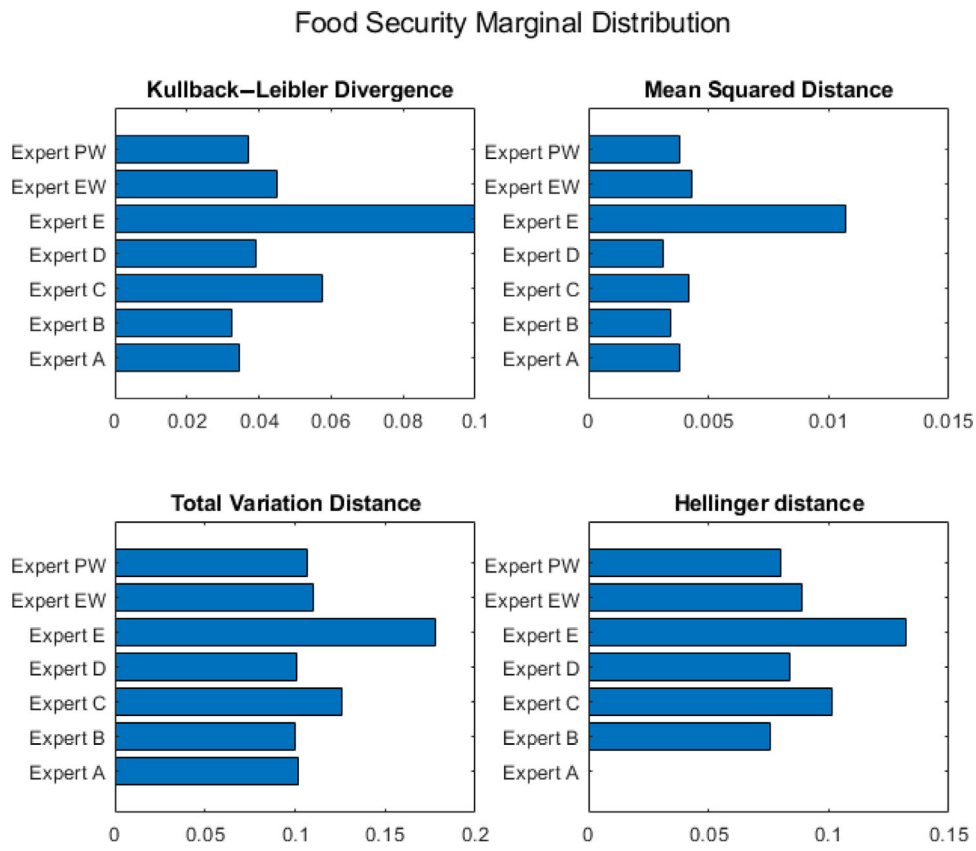

Fig 20. Food Security marginal probabilities and distance between expert and imputed values, as measured by the four selected measures. See Table A11.

We examine the result of these two different approaches below.[8]

2.3. Measures of Performance

In assessing the goodness of fit for the interpolated estimates, we use four measures of distance between probability distributions, namely, the mean squared deviation (MSD), the Kullback-Leibler divergence, total variation distance, and the Hellinger distance. The measures of performance that are usually used to verify the model against the truth, will, in this context, measure the distance between interpolated partial CPTs with fully elicited CPTs.

The MSD simply measures the average of the squares of the errors or deviations, that is, the difference between the estimator and what is estimated. MSD is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. If P is a vector of n predictions, and Q is the vector of observed values of the variable being predicted, then the MSD is computed as

$$\text{MSD} = \frac{1}{n}\sum_{i=1}^{n}(Q_i - P_i)^2.$$

This is similar to the Brier Score used to estimate difference between an expert's estimate and the reality, but in that case, the outcome is known. In this case, "InterBeta" is being compared to the expert's estimate. Values close to zero show good agreement.

The TVD is a distance measure for probability distributions and is an example of a statistical distance metric, and is sometimes called the statistical distance or variational distance. Computed as

$$TVD = \frac{1}{2}\sum_{i}(|P_i - Q_i|)$$

[8]Many "mixed" approaches are also possible. For example, one could aggregate just the best and worst cases, then interpolate each expert individually (using the parent weights that they provide) based on the group's best and worst cases. We do not examine these variations here.

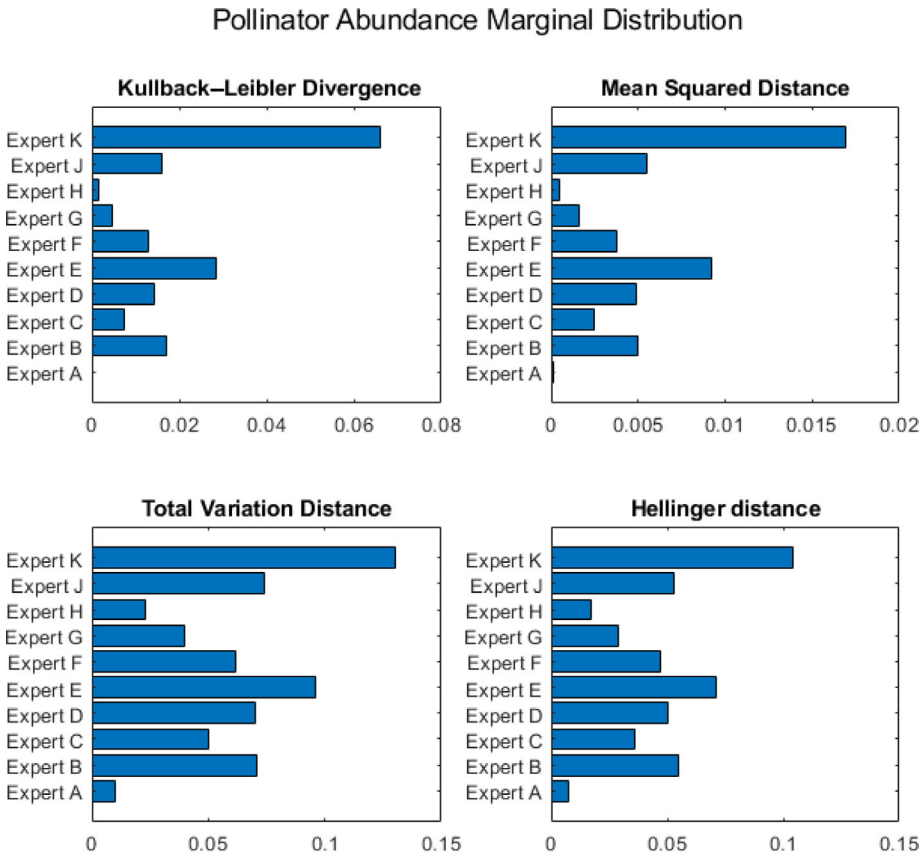

Fig 21. Pollinator abundance marginal probabilities and distance between expert and imputed values, as measured by the four selected measures. See Table A12.

Values close to zero show good agreement.

The Kullback–Leibler (KL) divergence from Q to P for discrete probability distributions P and Q is defined to be

$$D_{\mathrm{KL}}(P\|Q) = -\sum_i P(i)\,\log\frac{Q(i)}{P(i)}.$$

In other words, it is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P; smaller values show good agreement (MacKay, 2003, p. 34).

The Hellinger distance for two discrete probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ is defined as

$$H(P, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2},$$

which is directly related to the Euclidean norm of the difference of the square root vectors.

2.4. Cost Benefit Analysis

Experiments were conducted with each of the method options in InterBeta (see Table I), aiming to replicate as closely as possible the expert CPTs. For methods that require weights, we used simulated annealing to find the best fitting weights. The purpose of these experiments was to see how well InterBeta could do in principle, given the required inputs for each method. Essentially, we pretended that experts were able to provide weights that perfectly reflected their internal thinking in terms of the implied probabilities calculated for the CPTs. Of course, for Best and Worst, no pretence is necessary since no weights are elicited, and the best and worst case distributions would be provided by the expert in the same form as was used here.

The more sophisticated a variation of the InterBeta method is, the greater the elicitation burden. A cost-benefit analysis should enable us to establish if increasing this elicitation burden is worthwhile, that

is to say, is the increase in fidelity large enough. Of course, what large enough means depends on the application, the appetite of the experts to answer extra (and more complex) questions and many other factors.

3. RESULTS

3.1. Case Studies

Two SEJ exercises provide the "ground truth" to which the InterBeta methodology is compared. These elicitations were carried out for other research purposes, and in each case, the full CPTs were elicited using the IDEA protocol, giving a median and a 90% credible interval for each entry. InterBeta was provided with the minimal information it required for each of its modes of operation (as if we were only eliciting partial information from experts) and was used to estimate the values that have been withheld. Success is regarded as matching closely the values provided by each individual expert, or by the aggregated values obtained from the expert group. The aim is to show how accurate InterBeta is likely to be under various partial elicitation scenarios, and to provide guidance on which of the modes of operation might be optimal under different circumstances. The two SEJ exercises were part of a pollinator abundance model and a household food security model. The pollinator abundance model required a simpler elicitation of CPTs from a BN consisting of three binary parents and a binary child. Eliciting the entire CPT took two days in all, a one-day elicitation workshop and some follow-up online meetings to gather second-round estimates and conduct a calibration exercise. The household food security BN model was more complex, with four possible categories in the child node, and two binary and one ternary parent. Again, this took two days in all for elicitation of the complete CPT—a one-day elicitation workshop, which included the calibration questions, followed by an online discussion plus follow-up with participants who had been unable to attend the day at short notice.

3.1.1. Household Food Security in Victoria, Australia (Food)

In this experiment, a structured expert elicitation, using the IDEA protocol (Hanea et al., 2016), was undertaken with five domain experts in household food security (Barons, & Kleve, 2020). Experts were asked to estimate the proportions of 100 families in the state of Victoria, Australia, that would be in each of four categories of food security (High, Marginal, Low, Very Low) given various levels of physical access, equivalized disposable income and food prices. The responses are the median number out of 100 households who would be in each of the four categories; question 1 is the median number in high food security, question 2 in marginal food security, question 3 in low food security, and question 4 in very low food security. There were 12 scenarios, representing differing combinations of these three parent states, and four possible levels of food security in the target child node making 48 natural frequencies (probabilities) to estimate, plus an upper and lower plausible bound for each. There were, additionally, 20 calibration questions.

The "best case" scenario and "worst case" scenario have very natural interpretations here: good physical access, high equivalized disposable income, and low food prices (high food availability) should naturally lead to high rates of household food security while poor physical access, low equivalized disposable income, and high food prices (low food availability) should naturally lead to high rates of household food *in*security. We provide the natural frequencies from each of the experts, normalized to [0,1] and the equally weighted and performance-weighted combination of experts' estimates to provide explicit multinomials, required by InterBeta to fit a Beta distribution, and used to calculate missing values.

3.1.2. Pollinator Abundance (Bees)

In this experiment, a structured expert elicitation, using the IDEA protocol (Hanea et al., 2016), was undertaken with 10 domain experts in pollination and pollinators (Barons et al., 2018). Experts were asked to estimate the probability of *good* abundance of honey bees, wild bees, and other insect pollinators given all combinations of possible states of weather, the environment, and disease pressure. The probability of poor abundance was taken as 1 − Good Abundance. Evidence for the link between disease and honey bees is strong, but relatively incomplete for other bees and other pollinators. For this reason, we did not ask the experts to estimate the effects of disease on other bees and other pollinators. There were eight scenarios for honey bees and four

for each of the other bees and pollinators, making 16 probabilities to estimate, plus an upper and lower plausible bound for each.

The "best" case for this example is when Varroa (disease) control is good, weather is average (normal), and the environment is supportive. The "worst" case for this example is when Varroa control is poor, weather is unusual, and the environment is unsupportive. In this work, we used only the eight questions about honey bee abundance, since in the other cases, once the best and worst cases were defined, there were no others to interpolate.

3.2. Best and Worst

The simplest mode for InterBeta is when multinomials are elicited from experts for the "best" and "worst" cases and used to fit Beta distributions using the arithmetic mean.

3.2.1. Food

Tables A1–A6 in the Appendix show what the five experts gave for the intermediate cases, based on their answers to their views elicited on the proportions of 100 households in each of the four Food Security categories (High, Marginal, Low, Very Low) given the best and worst cases. Expert responses have been interleaved with the interpolator estimates for the same cases using the Best and Worst method for comparison. The interpolator had no examples from the human experts of situations where equivalized income is moderate. Fig. 5 shows the same information in graphical form. The red lines show the results the interpolator gave and the blue lines show the estimates from each of the experts themselves. Experts A–E are human experts; performance and equal weighted combinations are also shown. Each set of four questions relates to one specific scenario of income, access, and availability, which is indicated by the vertical lines in the table. The ragged shape of the distribution in the figures is because every number in the distribution is included serially; e.g., [0.93,0.04,0.02,0.01] followed immediately by [0.50,0.29,0.15,0.06], etc. However, using just the best and worst cases, InterBeta was able to capture this ragged distribution very well.

Taken over all questions, we used KL, MSD, TVD, and Hellinger distance between the expert results and the interpolator results to give a measure of overall accuracy (Fig.6).

For the Best and Worst method (with no user supplied weights), we see, in general, that the interpolator replicates the general shape, but tends to smooth the result, particularly away from the extremes, where the "true" values were given. When the bulk of the probability is in a single category, the interpolator significantly underestimates the probability here, and overestimates the probability in neighboring categories. This effect is particularly strong for most experts in questions 19–22, 37–40, 43–46, and 55–58.

In contrast, the difference was less for Q37–40 for Expert E, who had more of an economics background than others, and more for Q25–28 (moderate equivalized disposable income and good physical access when food availability is poor). However, as an overall performance, Expert A can be considered best, as measured by the Kulback–Leiblerscore, with a similar performance as the aggregations.

There is no clear difference (see Figs. 6 and 7) between imputing from the aggregated experts' judgments and aggregating the imputed values of the individuals experts' values. For equal weighting of the experts' medians, the aggregation of imputed values scores slightly better both in the Food example and in the Bee example below, but for performance weighted aggregation, there is a mixed picture, with the imputation of aggregated values tending to do slightly better.

3.2.2. Bees

Tables A10 and A11 show the interpolator estimates of what the 10 experts would have given for the intermediate cases, based on their answers to their views elicited on the probability of high abundance of honey bees given the best and worst cases.

3.3. Variations of Interpolation and Elicitation

Here we explore the Parent Weights, Parent State Weights, Row Weights, and Row Beta Parameters options within InterBeta.

Compared to the arithmetic mean and Best and Worst in the Food Security data, using parent weights gives a much closer fit, more of the peaks and troughs are captured in full, with similar performance whether fitting actual experts (A…E) or aggregations (EW and PW) (Fig. 9).

With parent state weights, the fit is closer still, with discrepancies seen mostly in Expert C. This

error is eliminated using the aggregated versions which fit very closely (Fig. 10).

Using the geometric mean with Best and Worst, we see the same pattern of smoothing away from the extremes as with the arithmetic mean, but if anything, the smoothing is greater using the geometric mean (Fig. 11).

As with arithmetic mean, geometric mean using parent weights brings a significant improvement in the performance, with the shape largely captured in full and the anomalies of Expert C eliminated through use of the aggregated values (Fig. 12). There does not seem to be a significant difference between equal weighted and performance weighted aggregation, which suggests in partially elicited CPTs interpolated using InterBeta, the burden on experts can be reduced by omitting calibration questions.

A further improvement in fit is seen using parent state weights with geometric mean (Fig. 13). The fit is particularly close with the aggregated values. Again, there is little to choose between these two, suggesting that calibration questions may not be necessary if partial CPTs are being elicited.

Comparing KL for the different imputations (see Fig. 14), we see that more inputs give greater accuracy, as expected, but also that in almost all cases, parent weights give a performance both significantly improved over Best and Worst and in line with Parent state weights, suggesting that the additional elicitation burden for the latter in not warranted in this case.

In the Bees SEJ, using Parent Weights and the arithmetic mean brings a noticeable improvement over Best and Worst (Fig. 15). Experts D, F, J, and K are particularly well captured.

The geometric mean captures the distributions using Best and Worst better than the arithmetic mean (Fig. 16).

Using Best and Worst and the geometric mean finds the best fit to the Bees experts (Fig. 17). The fit is slightly less closed in the case of Expert C and G, but is closed nevertheless.

Similarly to the Food example, in Fig. 18, we see an improvement in accuracy as more information is added along with a substantial improvement of Parent Weights over Best and Worst, but only a small additional improvement for Parent State Weights.

Comparison with Cain's method shows that, for the Bees SEJ, only 6 of the 16 values can be imputed, the remainder need to be provided to the algorithm (Fig. 19). The fit is reasonable in this case, but the elicitation burden is reduced by only about a third.

3.4. Effect on Marginal Probability

Recall Figs. 1, 3, and 4. The desired outcome of the elicitation was to quantify the CPTs of the BN to calculate the marginal probabilities of pollinator abundance and household food security. Our final interest is to investigate the difference between these marginal distributions and those derived from imputed CPTs using InterBeta with Best and Worst (i.e., no user defined weights) and the arithmetic mean. With uniform priors on the parents, we calculated the marginal probability distributions for household food security and bee abundance for each expert and for the aggregated distributions and compared these to the margins derived from the imputed values. In each case, we calculated the KL, MSE, TVD, and Hellinger distance

4. CONCLUSIONS

When SEJ is required to quantify models, one significant consideration is the burden on experts. This burden comes in the form of both time commitment and cognitive burden. Cognitive burden is expert-dependent and an elicitation that is burdensome to one expert may be easy to another, depending on the type of expertise and the quantities required. A large number of assessments are burdensome to all experts, even if easy, and InterBeta offers a way to reduce these burdens. InterBeta also provides the opportunity to enrich a model by making maximum use of available expert time.

In the two examples shown here, there was good agreement between the values the experts gave in the elicitation and the InterBeta values. The performance weighted aggregations performed well, but there is not a significant difference between interpolating the aggregated expert values and aggregating the interpolated values. In both food and bees examples, the parent state weights method in InterBeta produced estimates close to those provided by the experts. The elicitation requirement for this method is the best and worst cases and state weights that, as demonstrated, can be elicited using natural language. For the Food example, this represents a significant reduction in elicitation burden. In the Bees example, the burden is still reduced, but since this is a relatively small CPT, there is less saving. However, this CPT was reduced in order to make elicitations feasible in the time available. The addition of InterBeta to the SEJ toolbox offers the possibility of richer modeling. The widely used Cain's method performs well

on the (binary) bees case at the cost of much more expert burden—in fact, burden is decreased by only a third relative to eliciting the full CPTs, and this is a simple case. As more information is provided to InterBeta, the fidelity of the interpolation certainly improves, so the trade-off is between expert burden and accuracy, which is context-dependent. The elicitation burden for InterBeta Parent Weights (see Table 1) is the same as Cain's method. However, InterBeta has a range of options that allow trading off burden against fidelity, as well as fitting nonlinear relationships and better supporting dependencies when needed. For these variations, the elicitation burden can shrink (if only best and worst cases are requested) or grow if asking for weights for parent states, row weights, and Beta distribution parameters for each row. InterBeta is also able to scale up to arbitrarily large state spaces for both parents and children so long as the child continues to fit the shape of the (discretized) Beta distribution, with either a unimodal center of mass and dispersion, a uniform distribution, or a bimodal distribution with peaks at both ends. The elicitation burden in the case of Parent State Weights is linear in the number of states for each parent and the child—for example, for 10 parents with five states each, the number of weights required is $10 \times (5 - 1) = 40$, coupled with the best and worst case child distributions. In such a case, the limiting factor is more likely to be the BN inference itself rather than the time required from the expert. Practical (and well-engineered) causal graphs that are human-friendly are typically much sparser than this; hence, we might expect this method to work feasibly and well with networks that run to a hundred or more nodes.

The elicitation burden, however, does depend to a degree on the experts' areas of expertise. Many domain experts would be very comfortable with providing parameters for distributions, while others would find this way of thinking very unfamiliar and potentially confusing, which affects the reliability of the results they would give. One assessment that must be undertaken is the appropriateness of the methods for the specific experts invited to provide quantitative information for the problem at hand. For the modeling methods reviewed above, the likelihood method requires experts to be comfortable with probability distributions. The EBBN method requires giving rows for each of the child states and parent weights. Some experts may be more comfortable with this. The relative weights for the parent nodes and probability distributions for compatible parent configurations required by the weighted sum algorithm will prove challenging for experts from some domains. Similarly, providing an aggregation function for the RNM along with parent weights and parameter variance may prove to be a heavy burden for some domain experts and straightforward for others. The examination and editing of the CPTs or refinement of the interval scales, required in the RNM and the ACE methods, implies that the experts already have an idea of what these are and could, in principle, have provided the CPTs by hand. Of course, if the CPTs are very large, initializing them still reduces expert burden to some degree. Asking for qualitative descriptors and translating these into numerical values will reduce the burden for experts in fields where descriptive categorizations are the norm. However, translating these into numerical scores appropriately will require domain expertise as well as mathematical expertise. There is also an issue here that the qualitative information may be interpreted differently by different individuals. This can be addressed by agreeing on a numerical range within which each categorical descriptor lies. The complexity of a system model is often not fully known until domain experts are consulted, showing that the soft elicitation or model building phase must not be rushed or undertaken by nonexperts. This applies to modeling generally and not just to BNs where interpolation might be used. The most important quantity is the difference in the marginal distribution to the quantity of interest that the BN is estimating, and on which decision will be made.

The flow chart in Fig. 2 gives readers a logical way to understand the way InterBeta might be used to best effect in a given context. Compared to existing methods, InterBeta provides a flexible tool that allows control of expert burden and trade-off between burden and accuracy in a transparent single method. The examples given here provide a proof of concept and insight into the potential accuracy of InterBeta. Further work will be required to provide a theoretical basis for this.

ACKNOWLEDGMENTS

The authors would like to thank Prof Ann Nicholson of Department of Computer Science and Bayesian Intelligence, Monash University for helpful discussions. Research supported by EPSRC grant EP/K039628/1

APPENDIX A

Table A1. Five Experts Estimated the Number Out of 100 Families to Experience Each of Four Categories of Food Security Using the IDEA Protocol

Expert Estimate	Question 1	Question 2	Question 3	Question 4	Question 7	Question 8	Question 9	Question 10
Expert A	0.93	0.04	0.02	0.01	0.50	0.29	0.15	0.06
Expert A Imputed	0.93	0.04	0.02	0.01	0.70	0.18	0.09	0.03
Expert B	0.90	0.06	0.03	0.01	0.36	0.44	0.12	0.08
Expert B Imputed	0.90	0.06	0.03	0.01	0.66	0.19	0.10	0.04
Expert C	0.95	0.03	0.02	0.01	0.20	0.60	0.15	0.05
Expert C Imputed	0.95	0.03	0.02	0.01	0.60	0.19	0.13	0.08
Expert D	0.85	0.10	0.03	0.02	0.35	0.47	0.15	0.03
Expert D Imputed	0.85	0.09	0.04	0.02	0.65	0.20	0.11	0.04
Expert E	0.97	0.01	0.01	0.01	0.85	0.05	0.05	0.05
Expert E Imputed	0.97	0.01	0.01	0.01	0.71	0.11	0.08	0.10
Equal Weight	0.93	0.04	0.02	0.01	0.42	0.39	0.14	0.06
Equal Weight Imputed	0.93	0.04	0.02	0.01	0.66	0.18	0.11	0.05
Performance Weight	0.88	0.08	0.03	0.01	0.35	0.45	0.14	0.07
Performance Weight Imp	0.88	0.07	0.03	0.01	0.65	0.20	0.11	0.04

Expert Estimate	Question 13	Question 14	Question 15	Question 16	Question 19	Question 20	Question 21	Question 22
Expert A	0.28	0.27	0.37	0.08	0.90	0.05	0.04	0.01
Expert A Imputed	0.51	0.27	0.16	0.06	0.51	0.27	0.16	0.06
Expert B	0.25	0.25	0.40	0.10	0.85	0.10	0.03	0.02
Expert B Imputed	0.48	0.27	0.17	0.07	0.48	0.27	0.17	0.07
Expert C	0.13	0.30	0.51	0.06	0.75	0.12	0.10	0.03
Expert C Imputed	0.39	0.27	0.21	0.13	0.39	0.27	0.21	0.13
Expert D	0.15	0.35	0.45	0.05	0.90	0.08	0.01	0.01
Expert D Imputed	0.48	0.27	0.17	0.08	0.48	0.27	0.17	0.08
Expert E	0.20	0.10	0.10	0.60	0.97	0.01	0.01	0.01
Expert E Imputed	0.50	0.17	0.14	0.19	0.50	0.17	0.14	0.19
Equal Weight	0.22	0.28	0.40	0.10	0.87	0.08	0.04	0.02
Equal Weight Imputed	0.46	0.27	0.18	0.09	0.46	0.27	0.18	0.09
Performance Weight	0.22	0.29	0.41	0.08	0.86	0.09	0.03	0.02
Performance Weight Imp	0.48	0.27	0.17	0.08	0.48	0.27	0.17	0.08

Expert Estimate	Question 25	Question 26	Question 27	Question 28	Question 31	Question 32	Question 33	Question 34
Expert A	0.27	0.45	0.18	0.09	0.21	0.24	0.42	0.13
Expert A Imputed	0.37	0.32	0.22	0.09	0.26	0.34	0.27	0.12
Expert B	0.45	0.40	0.10	0.05	0.20	0.25	0.42	0.13
Expert B Imputed	0.35	0.31	0.23	0.11	0.25	0.33	0.28	0.14
Expert C	0.30	0.60	0.06	0.04	0.25	0.25	0.30	0.20
Expert C Imputed	0.26	0.30	0.27	0.17	0.17	0.30	0.32	0.21
Expert D	0.30	0.52	0.15	0.03	0.20	0.22	0.33	0.25
Expert D Imputed	0.35	0.31	0.23	0.11	0.26	0.32	0.28	0.14
Expert E	0.60	0.25	0.05	0.10	0.20	0.10	0.10	0.60
Expert E Imputed	0.34	0.18	0.18	0.30	0.23	0.18	0.19	0.40
Equal Weight	0.35	0.47	0.12	0.06	0.22	0.24	0.35	0.19
Equal Weight Imputed	0.32	0.30	0.24	0.13	0.23	0.32	0.29	0.17
Performance Weight	0.41	0.43	0.12	0.05	0.20	0.24	0.38	0.18
Performance Weight Imp	0.34	0.31	0.23	0.11	0.25	0.33	0.28	0.14

(*Continued*)

Table A1. (Continued)

Expert Estimate	Question 37	Question 38	Question 39	Question 40	Question 43	Question 44	Question 45	Question 46
Expert A	0.88	0.07	0.03	0.02	0.68	0.19	0.08	0.05
Expert A Imputed	0.51	0.27	0.16	0.06	0.37	0.32	0.22	0.09
Expert B	0.85	0.08	0.05	0.02	0.60	0.25	0.10	0.05
Expert B Imputed	0.48	0.27	0.17	0.07	0.35	0.31	0.23	0.11
Expert C	0.80	0.10	0.06	0.04	0.30	0.60	0.08	0.02
Expert C Imputed	0.39	0.27	0.21	0.13	0.26	0.30	0.27	0.17
Expert D	0.85	0.10	0.03	0.02	0.60	0.30	0.07	0.03
Expert D Imputed	0.48	0.27	0.17	0.08	0.35	0.31	0.23	0.11
Expert E	0.80	0.10	0.08	0.02	0.70	0.10	0.10	0.10
Expert E Imputed	0.50	0.17	0.14	0.19	0.34	0.18	0.18	0.30
Equal Weight	0.84	0.09	0.05	0.02	0.58	0.29	0.08	0.05
Equal Weight Imputed	0.46	0.27	0.18	0.09	0.32	0.30	0.24	0.13
Performance Weight	0.84	0.09	0.05	0.02	0.58	0.29	0.08	0.04
Performance Weight Imp	0.48	0.27	0.17	0.08	0.34	0.31	0.23	0.11

Expert Estimate	Question 49	Question 50	Question 51	Question 52	Question 55	Question 56	Question 57	Question 58
Expert A	0.33	0.25	0.33	0.10	0.79	0.15	0.04	0.02
Expert A Imputed	0.26	0.34	0.27	0.12	0.26	0.34	0.27	0.12
Expert B	0.20	0.25	0.43	0.12	0.80	0.11	0.06	0.03
Expert B Imputed	0.25	0.33	0.28	0.14	0.25	0.33	0.28	0.14
Expert C	0.20	0.20	0.30	0.30	0.40	0.35	0.20	0.05
Expert C Imputed	0.17	0.30	0.32	0.21	0.17	0.30	0.32	0.21
Expert D	0.25	0.35	0.25	0.15	0.80	0.15	0.03	0.02
Expert D Imputed	0.26	0.32	0.28	0.14	0.26	0.32	0.28	0.14
Expert E	0.20	0.10	0.10	0.60	0.80	0.10	0.08	0.02
Expert E Imputed	0.23	0.18	0.19	0.40	0.23	0.18	0.19	0.40
Equal Weight	0.26	0.24	0.33	0.18	0.75	0.16	0.07	0.03
Equal Weight Imputed	0.23	0.32	0.29	0.17	0.23	0.32	0.29	0.17
Performance Weight	0.23	0.27	0.36	0.14	0.79	0.13	0.05	0.03
Performance Weight Imp	0.25	0.33	0.28	0.14	0.25	0.33	0.28	0.14

Expert Estimate	Question 61	Question 62	Question 63	Question 64	Question 65	Question 66	Question 67	Question 68
Expert A	0.43	0.33	0.19	0.05	0.16	0.22	0.46	0.15
Expert A Imputed	0.19	0.34	0.32	0.15	0.13	0.33	0.36	0.18
Expert B	0.45	0.40	0.10	0.05	0.15	0.22	0.45	0.18
Expert B Imputed	0.17	0.33	0.33	0.17	0.12	0.31	0.36	0.20
Expert C	0.30	0.50	0.10	0.10	0.10	0.20	0.45	0.25
Expert C Imputed	0.12	0.29	0.35	0.24	0.08	0.27	0.38	0.27
Expert D	0.50	0.40	0.07	0.03	0.15	0.25	0.40	0.20
Expert D Imputed	0.19	0.32	0.32	0.18	0.13	0.30	0.35	0.21
Expert E	0.60	0.20	0.10	0.10	0.10	0.10	0.20	0.60
Expert E Imputed	0.15	0.15	0.19	0.50	0.09	0.13	0.18	0.60
Equal Weight	0.46	0.38	0.10	0.06	0.13	0.22	0.44	0.21
Equal Weight Imputed	0.16	0.31	0.33	0.20	0.11	0.30	0.36	0.23
Performance Weight	0.47	0.40	0.08	0.05	0.15	0.23	0.43	0.19
Performance Weight Imp	0.18	0.32	0.32	0.18	0.12	0.31	0.36	0.21

Note: Mathematical aggregation was performed with equal and performance weighting. Imputed values used Best and Worst (i.e., the smallest number of parameters of all methods) and arithmetic mean.

Table A2. Food Example Using the Mean Squared Deviation Measure: PW Parent Weights, SW Parent State Weights, RW Row Weights, RP Row Beta Parameters

Expert	PW-MSD	SW-MSD	RW-MSD	RP-MSD
Expert A	0.01	0.005	0.002	0.001
Expert B	0.006	0.004	0.003	0.002
Expert C	0.02	0.02	0.01	0.003
Expert D	0.007	0.006	0.004	0.0009
Expert E	0.009	0.003	0.002	0.0003
Equal Weighted Experts	0.006	0.004	0.003	0.001
Performance Weighted Experts	0.005	0.004	0.003	0.002

Table A3. Food Example Using the Total Variation Distance Measure: PW Parent Weights, SW Parent State Weights, RW Row Weights, RP Row Beta Parameters

Expert	PW-TVD	SW-TVD	RW-TVD	RP-TVD
Expert A	0.11	0.10	0.06	0.05
Expert B	0.12	0.10	0.08	0.07
Expert C	0.19	0.18	0.15	0.07
Expert D	0.12	0.11	0.09	0.04
Expert E	0.13	0.09	0.06	0.02
Equal Weighted Experts	0.11	0.09	0.07	0.05
Performance Weighted Experts	0.11	0.09	0.07	0.06

Table A4. Food Example Using the Hellinger Distance Measure: PW Parent Weights, SW Parent State Weights, RW Row Weights, RP Row Beta Parameters

Expert	PW-Hellinger	SW-Hellinger	RW-Hellinger	RP-Hellinger
Expert A	0.10	0.09	0.06	0.05
Expert B	0.10	0.08	0.07	0.06
Expert C	0.16	0.15	0.13	0.06
Expert D	0.12	0.11	0.09	0.06
Expert E	0.12	0.09	0.07	0.03
Expert EW	0.10	0.08	0.06	0.05
Expert PW	0.09	0.08	0.07	0.05

Table A5. Ten Experts Estimated the Probability of Good Abundance of Honey Bees, Other Bees, and Other Pollinators Using the IDEA Protocol

Expert Estimate	Question 1	Question 1a	Question 2	Question 2a	Question 3	Question 3a	Question 4	Question 4a
Expert A	0.70	0.30	0.20	0.80	0.60	0.40	0.25	0.75
Expert A Imputed	0.70	0.30	0.47	0.53	0.47	0.53	0.28	0.72
Expert B	0.85	0.15	0.10	0.90	0.40	0.60	0.10	0.90
Expert B Imputed	0.85	0.15	0.43	0.57	0.43	0.57	0.16	0.84
Expert C	0.75	0.25	0.25	0.75	0.40	0.60	0.30	0.70
Expert C Imputed	0.75	0.25	0.54	0.46	0.54	0.46	0.35	0.65
Expert D	0.80	0.20	0.40	0.60	0.70	0.30	0.30	0.70
Expert D Imputed	0.80	0.20	0.49	0.51	0.49	0.51	0.24	0.76
Expert E	0.80	0.20	0.25	0.75	0.40	0.60	0.30	0.70
Expert E Imputed	0.80	0.20	0.50	0.50	0.50	0.50	0.25	0.75
Expert F	0.75	0.25	0.35	0.65	0.60	0.40	0.25	0.75
Expert F Imputed	0.75	0.25	0.44	0.56	0.44	0.56	0.22	0.78
Expert G	0.80	0.20	0.25	0.75	0.50	0.50	0.20	0.80
Expert G Imputed	0.80	0.20	0.56	0.44	0.56	0.44	0.32	0.68
Expert H	0.75	0.25	0.35	0.65	0.45	0.55	0.30	0.70
Expert H Imputed	0.75	0.25	0.51	0.49	0.51	0.49	0.29	0.71
Expert J	0.80	0.20	0.20	0.80	0.70	0.30	0.15	0.85
Expert J Imputed	0.80	0.20	0.43	0.57	0.43	0.57	0.19	0.81
Expert K	0.70	0.30	0.35	0.65	0.45	0.55	0.20	0.80
Expert K Imputed	0.70	0.30	0.17	0.83	0.17	0.83	0.06	0.94
Equal Weight	0.77	0.23	0.27	0.73	0.52	0.48	0.24	0.76
Equal Weight Imputed	0.89	0.11	0.76	0.24	0.76	0.24	0.53	0.47

Expert Estimate	Question 5	Question 5a	Question 6	Question 6a	Question 7	Question 7a	Question 8	Question 8a
Expert A	0.20	0.80	0.15	0.85	0.15	0.85	0.15	0.85
Expert A Imputed	0.47	0.53	0.28	0.72	0.28	0.72	0.15	0.85
Expert B	0.50	0.50	0.05	0.95	0.30	0.70	0.05	0.95
Expert B Imputed	0.43	0.57	0.16	0.84	0.16	0.84	0.05	0.95
Expert C	0.30	0.70	0.20	0.80	0.50	0.50	0.20	0.80
Expert C Imputed	0.54	0.46	0.35	0.65	0.35	0.65	0.20	0.80
Expert D	0.40	0.60	0.15	0.85	0.40	0.60	0.10	0.90
Expert D Imputed	0.49	0.51	0.24	0.76	0.24	0.76	0.10	0.90
Expert E	0.40	0.60	0.20	0.80	0.20	0.80	0.10	0.90
Expert E Imputed	0.50	0.50	0.25	0.75	0.25	0.75	0.10	0.90
Expert F	0.40	0.60	0.10	0.90	0.30	0.70	0.10	0.90
Expert F Imputed	0.44	0.56	0.22	0.78	0.22	0.78	0.10	0.90
Expert G	0.45	0.55	0.20	0.80	0.45	0.55	0.15	0.85
Expert G Imputed	0.56	0.44	0.32	0.68	0.32	0.68	0.15	0.85
Expert H	0.35	0.65	0.15	0.85	0.40	0.60	0.15	0.85
Expert H Imputed	0.51	0.491	0.291	0.71	0.291	0.71	0.15	0.85
Expert J	0.60	0.40	0.10	0.90	0.50	0.50	0.07	0.93
Expert J Imputed	0.43	0.57	0.19	0.81	0.19	0.81	0.07	0.93
Expert K	0.20	0.80	0.10	0.90	0.10	0.90	0.02	0.98
Expert K Imputed	0.17	0.83	0.06	0.94	0.06	0.94	0.02	0.98
Equal Weight	0.38	0.62	0.14	0.86	0.33	0.67	0.11	0.89
Equal Weight Imputed	0.76	0.24	0.53	0.47	0.53	0.47	0.23	0.77

Note: Performance weights did not yield any benefit, so the mathematical aggregation was performed with equal weighting only. Imputed values were derived from InterBeta using Best and Worst (i.e., the smallest number of parameters of all methods) and arithmetic mean.

Table A6. Bees Example Using Mean Squared Deviation Measure: PW Arithmetic Mean, Parent Weights; EW Geometric Mean, Best and Worst; PWG Geometric Mean, Parent Weights; Cain Cain's Method

Expert	PW-MSD	EW-MSD	PWG-MSD	Cain-MSD
Expert A	0.011	0.015	0.000	0.001
Expert B	0.006	0.021	0.001	0.001
Expert C	0.023	0.013	0.008	0.010
Expert D	0.001	0.029	0.002	0.001
Expert E	0.009	0.004	0.002	0.003
Expert F	0.0003	0.019	0.002	0.002
Expert G	0.011	0.010	0.002	0.002
Expert H	0.007	0.006	0.003	0.004
Expert J	0.001	0.060	0.001	0.001
Expert K	0.005	0.032	0.001	0.000
Equal Weighted Experts	0.003	0.010	0.001	0.001
Average of generated CPTs	0.002	0.011	0.001	0.000

Table A7. Bees Example Using Total Variation Distance Measure: PW Arithmetic Mean, Parent Weights; EW Geometric Mean, Best and Worst; PWG Geometric Mean, Parent Weights; Cain Cain's Method

Expert	PW-TVD	EW-TVD	EW-TVD	Cain-TVD
Expert A	0.08	0.092	0.092	0.013
Expert B	0.05	0.106	0.106	0.014
Expert C	0.11	0.084	0.084	0.044
Expert D	0.019	0.120	0.120	0.022
Expert E	0.071	0.055	0.055	0.025
Expert F	0.012	0.103	0.103	0.024
Expert G	0.078	0.079	0.079	0.018
Expert H	0.064	0.058	0.058	0.033
Expert J	0.019	0.163	0.163	0.015
Expert K	0.05	0.132	0.132	0.009
Equal Weighted Experts	0.037	0.074	0.074	0.015
Average of generated CPTs	0.034	0.076	0.076	0.013

Table A8. Bees Example Using Hellinger Distance Measure: PW Arithmetic Mean, Parent Weights; EW Geometric Mean, Best and Worst; PWG Geometric Mean, Parent Weights; Cain Cain's Method

Expert	PW-Hellinger	EW-Hellinger	PWG-Hellinger	Cain-Hellinger
Expert A	0.0651	0.073	0.014	0.012
Expert B	0.047	0.093	0.021	0.015
Expert C	0.078	0.064	0.046	0.034
Expert D	0.015	0.093	0.031	0.018
Expert E	0.053	0.045	0.029	0.022
Expert F	0.01	0.083	0.033	0.021
Expert G	0.058	0.061	0.023	0.013
Expert H	0.048	0.046	0.038	0.027
Expert J	0.016	0.126	0.022	0.015
Expert K	0.057	0.134	0.020	0.009
Equal Weighted Experts	0.028	0.059	0.022	0.013
Average of generated CPTs	0.026	0.061	0.020	0.011

Table A9. Food Security SEJ: Performance Measures for Best and Worst Interpolation of Food Security SEJ CPTs

Expert	KL	MSD	TVD	Hellinger
Expert A	0.18	0.03	0.23	0.19
Expert B	0.20	0.02	0.25	0.20
Expert C	0.23	0.03	0.27	0.22
Expert D	0.23	0.03	0.26	0.21
Expert E	0.26	0.04	0.27	0.22
Expert EW	0.18	0.03	0.23	0.19
Expert PW	0.19	0.03	0.24	0.19

Note: "Equal weighted comb" is the mean of the imputed values compared with the mean of the experts' values, and equivalently for "Performance weighted comb." Performance weight is the results from InterBeta treating this as an "expert." See Figure 5.

Table A10. Pollinator Abundance SEJ

Expert Estimate	KL	MSD	TVD	Hellinger
Expert A	0.46	0.03	0.97	0.21
Expert B	0.41	0.02	0.74	0.17
Expert C	0.44	0.03	1.02	0.09
Expert D	0.22	0.01	0.70	0.02
Expert E	0.19	0.01	0.61	0.06
Expert F	0.14	0.01	0.52	0.04
Expert G	0.35	0.02	0.87	0.09
Expert H	0.19	0.01	0.62	0.06
Expert J	0.61	0.03	1.10	0.08
Expert K	0.45	0.02	0.71	0.30
Equal Weight	0.16	0.01	0.54	0.06
Equal weighted comb	0.14	0.01	0.52	0.05

Note: There was no performance weighting in this data set since the measures of performance were unable to justify performance weighting. Equal weighted comb is the mean of the imputed values compared with the mean of the experts' values (Equal Weight). See Fig. 8.

Table A11. Food Marginal Probabilities and Distance between Expert and Imputed Values

	KL	MSD	TVD	Hellinger
Expert A	0.0345	0.0038	0.1020	0.0000
Expert B	0.0326	0.0034	0.1000	0.0754
Expert C	0.0575	0.0042	0.1260	0.1013
Expert D	0.0390	0.0031	0.1010	0.0837
Expert E	0.0997	0.0107	0.1780	0.1322
Expert EW	0.0450	0.0043	0.1100	0.0890
Expert PW	0.0371	0.0038	0.1063	0.0803

Table A12. Bees Marginal Probabilities and Distance between Expert and Imputed Values

	KL	MSD	TVD	Hellinger
Expert A	0.0003	0.0001	0.0100	0.0075
Expert B	0.0172	0.0050	0.0710	0.0551
Expert C	0.0075	0.0025	0.0500	0.0362
Expert D	0.0144	0.0049	0.0700	0.0500
Expert E	0.0284	0.0092	0.0960	0.0708
Expert F	0.0129	0.0038	0.0620	0.0468
Expert G	0.0048	0.0016	0.0400	0.0290
Expert H	0.0017	0.0005	0.0230	0.0173
Expert J	0.0162	0.0055	0.0740	0.0532
Expert K	0.0662	0.0169	0.1300	0.1042

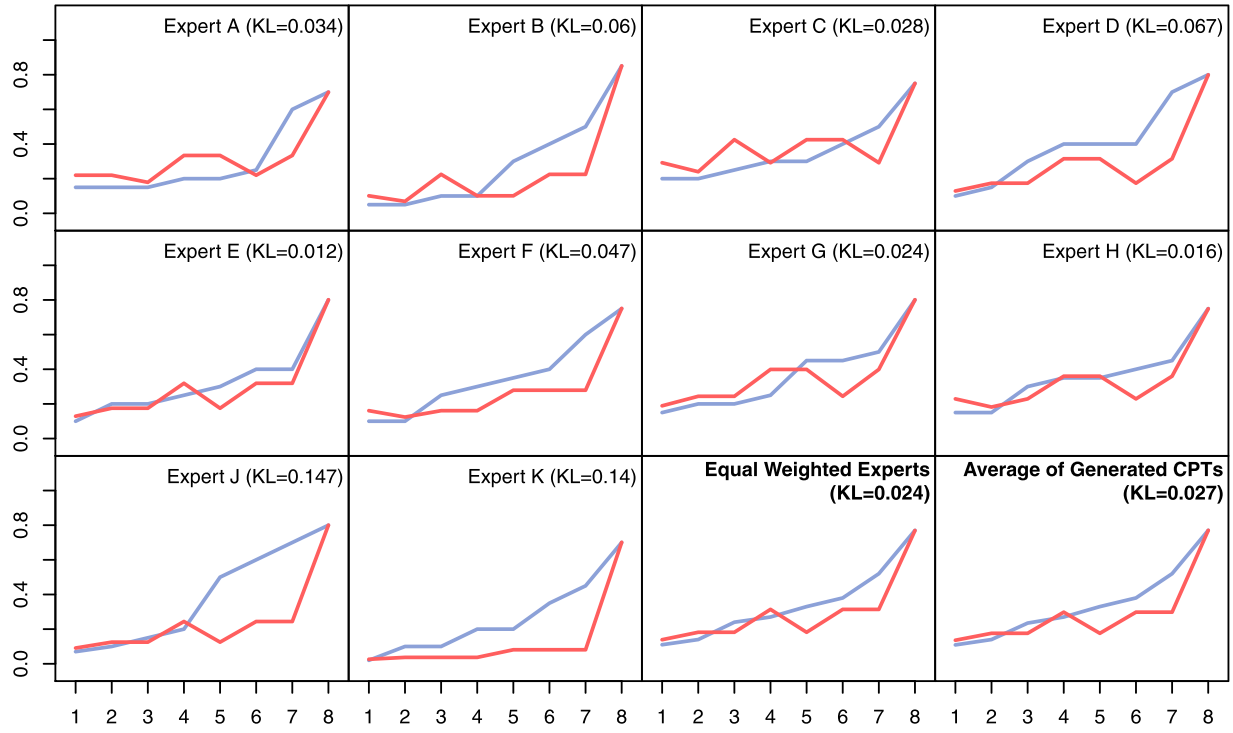

Fig A1. Bee geometric best and worst.

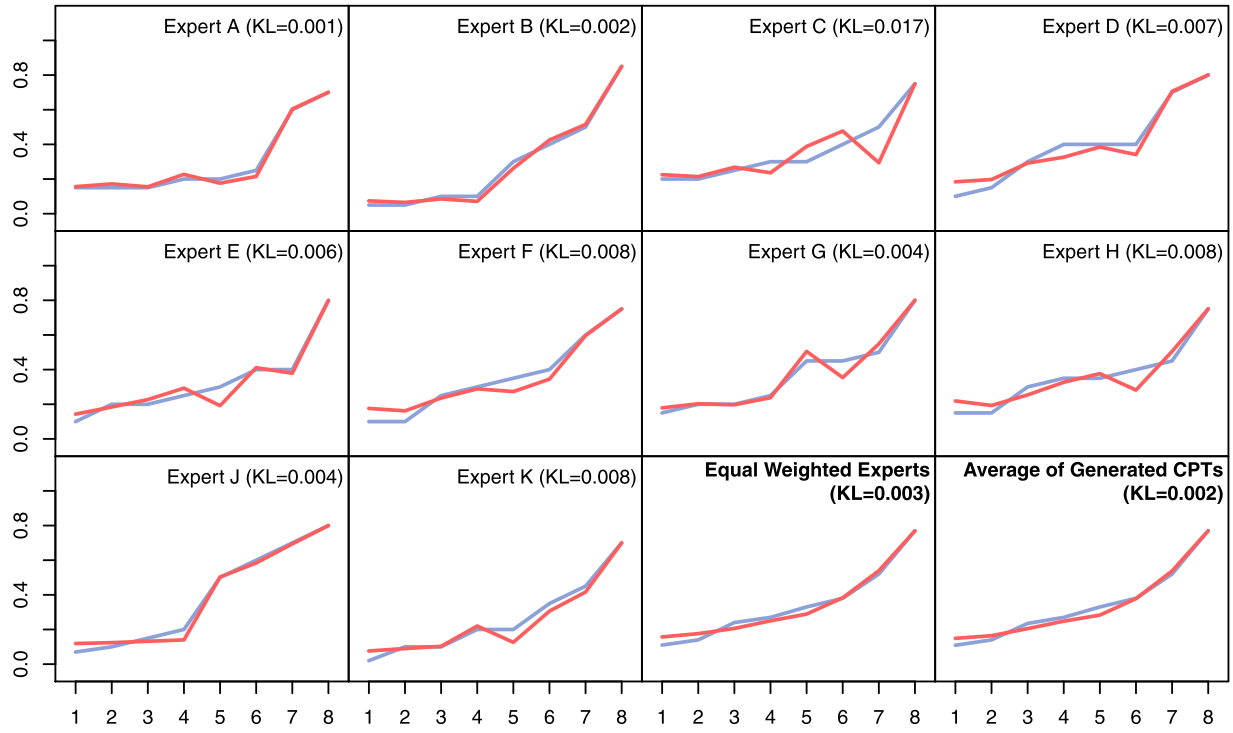

Fig A2. Bee geometric parent weights.

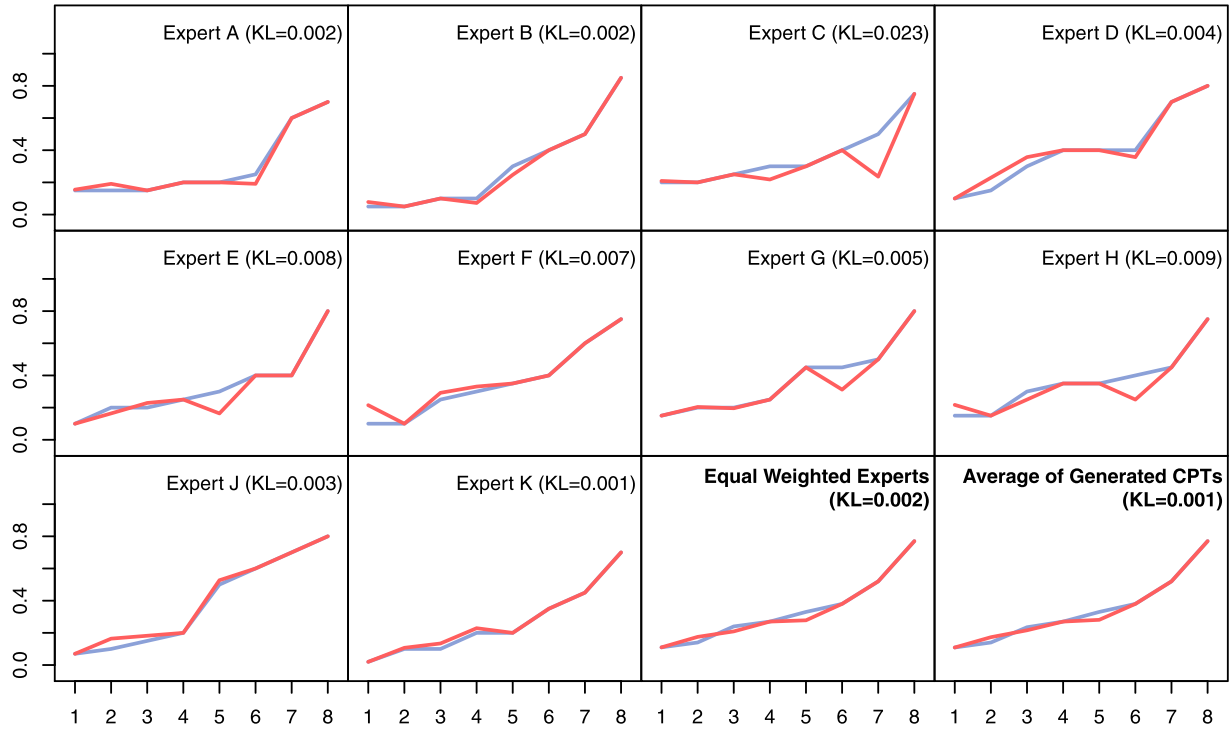

Fig A3. Bee cain.

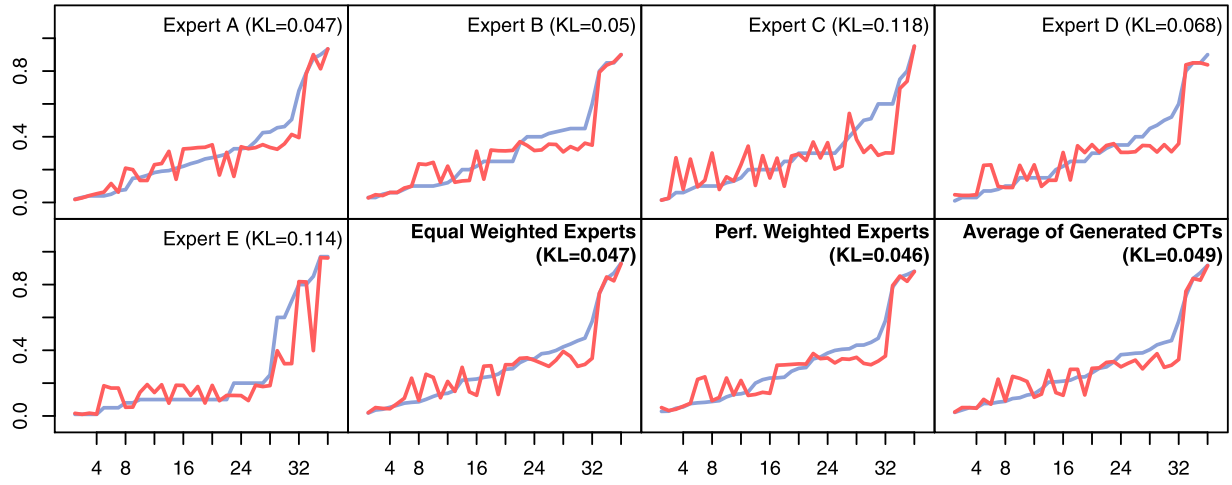

Fig A4. Food arithmetic parent weights.

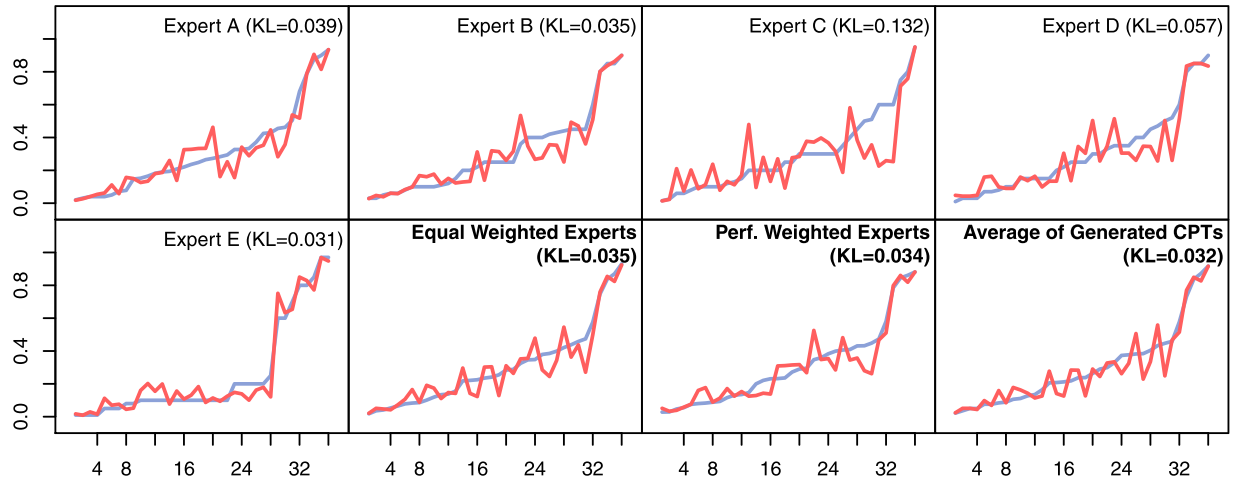

Fig A5. Food arithmetic parent state weights.

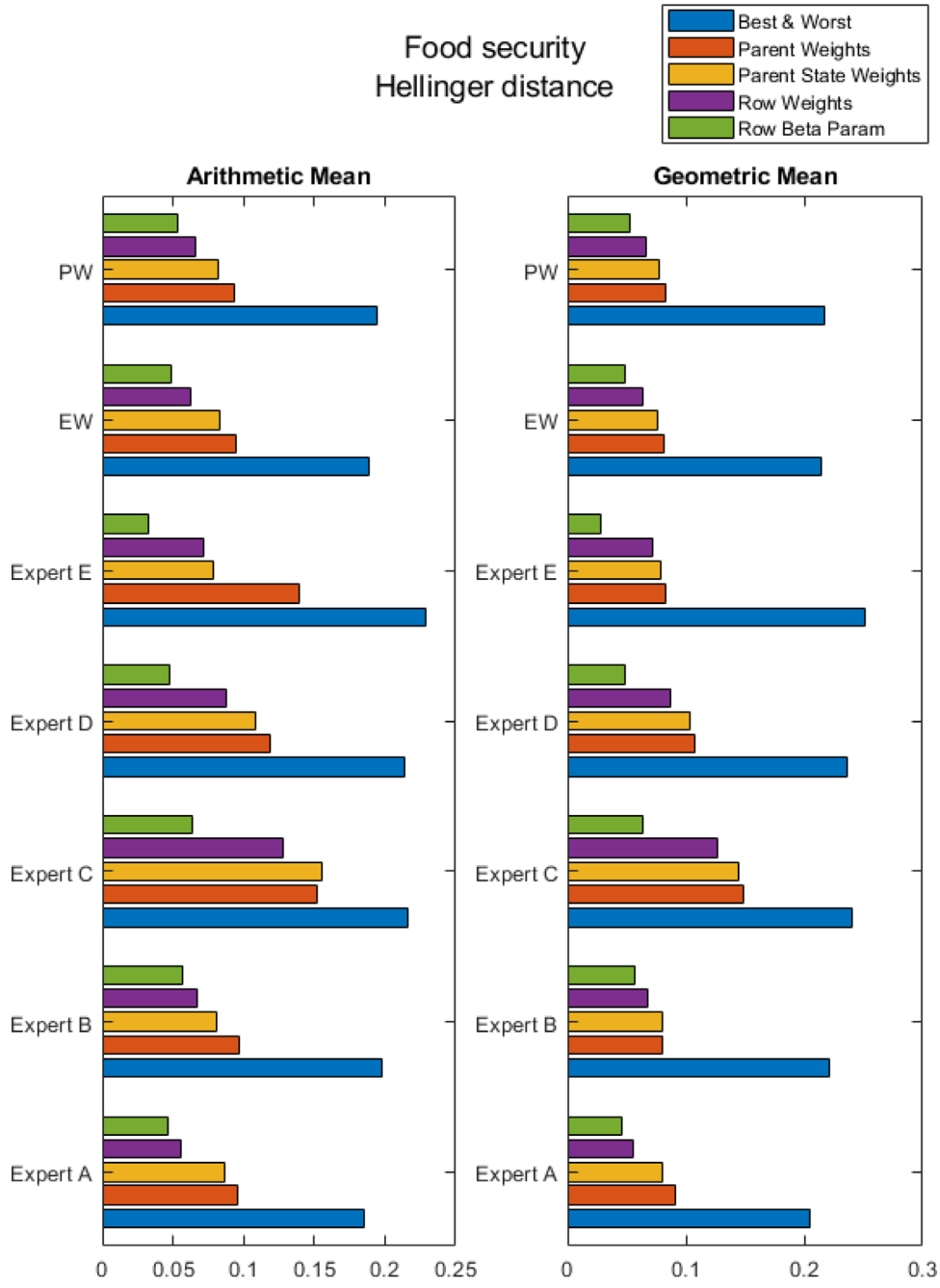

Fig A6. Food—Hellinger—Arithmetic and Geometric Mean—Comparison.

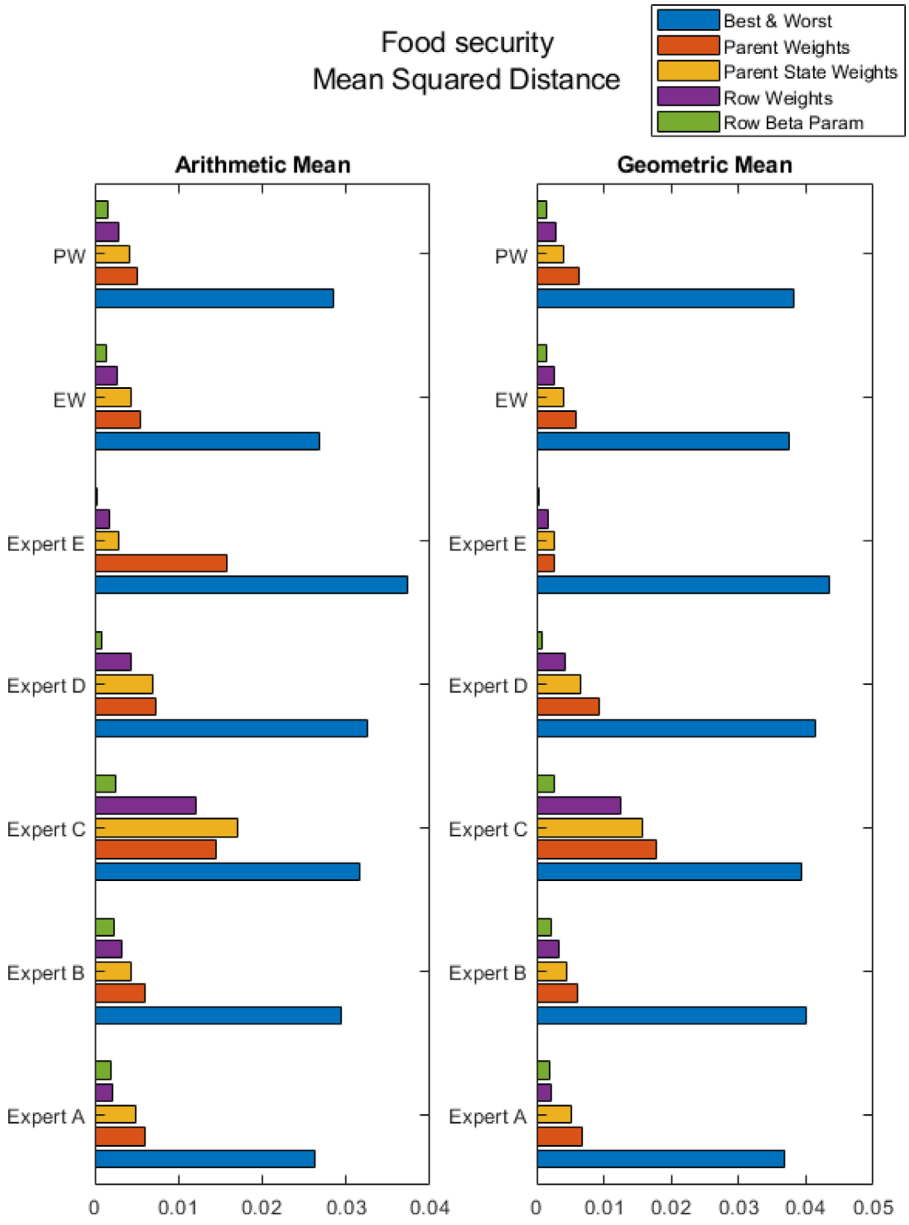

Fig A7. Food—MSD—Arithmetic and Geometric Mean—Comparison.

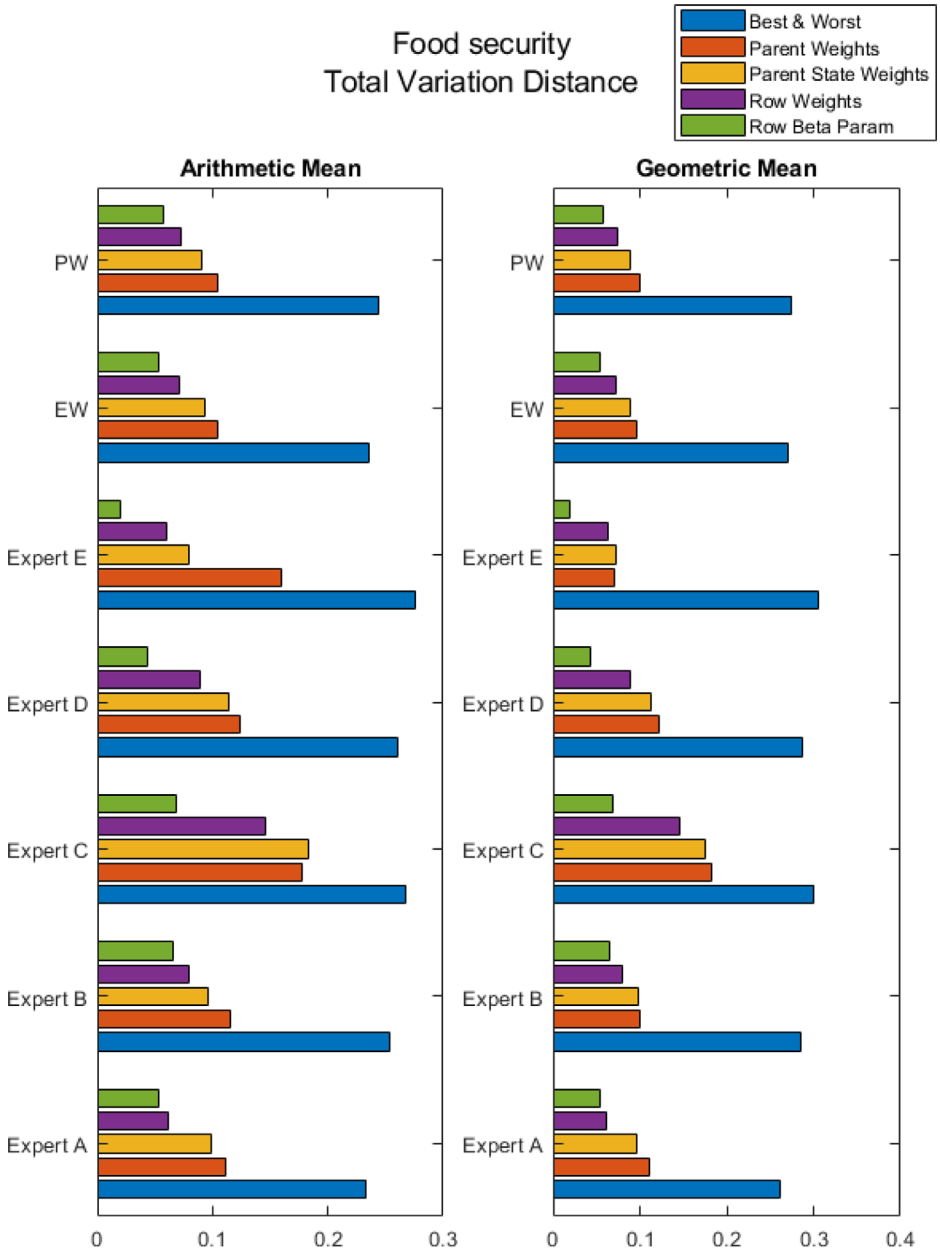

Fig A8. Food—TVD—Arithmetic and Geometric Mean—Comparison.

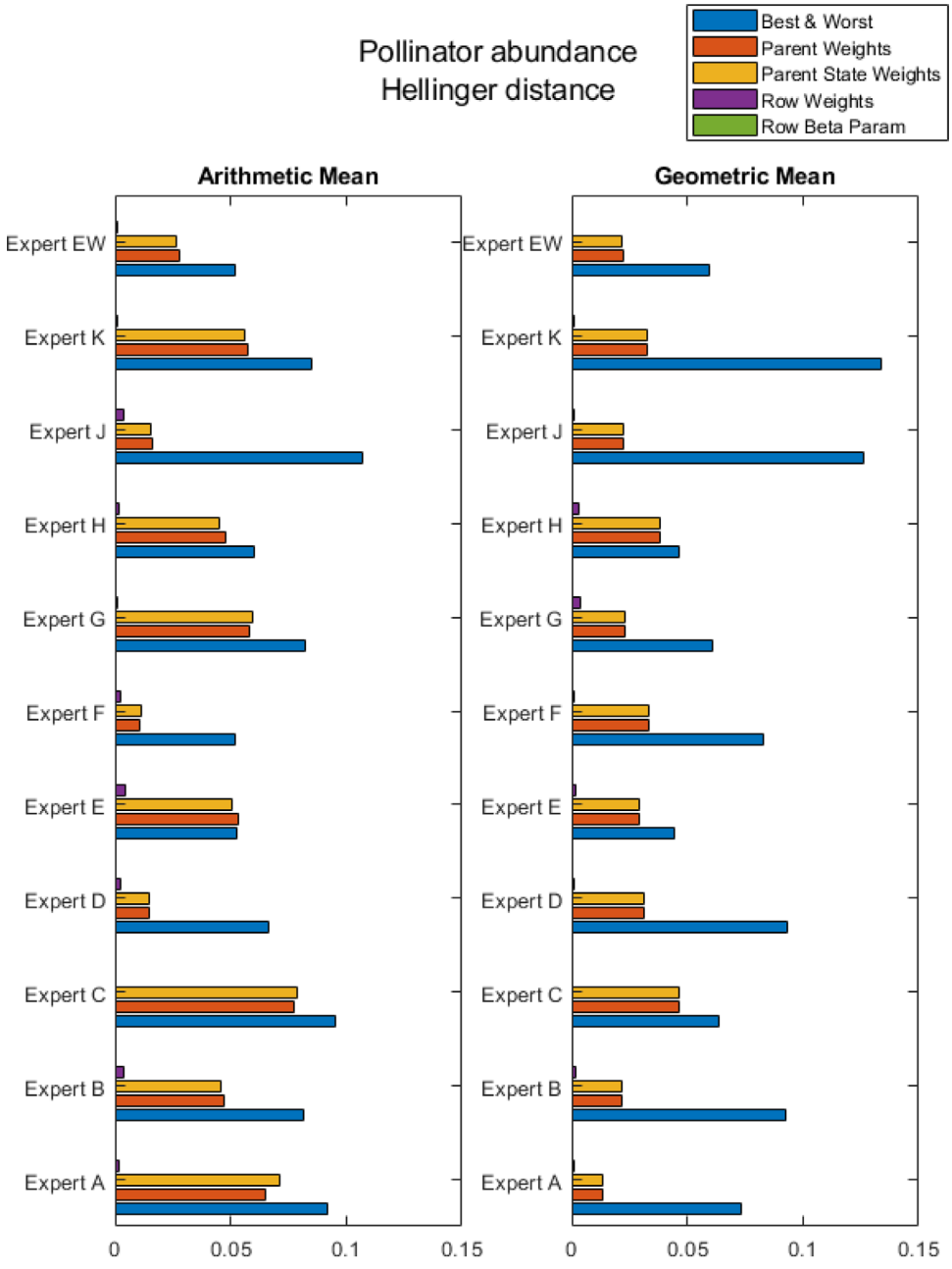

Fig A9. Bees—Hellinger—Arithmetic and Geometric Mean—Comparison.

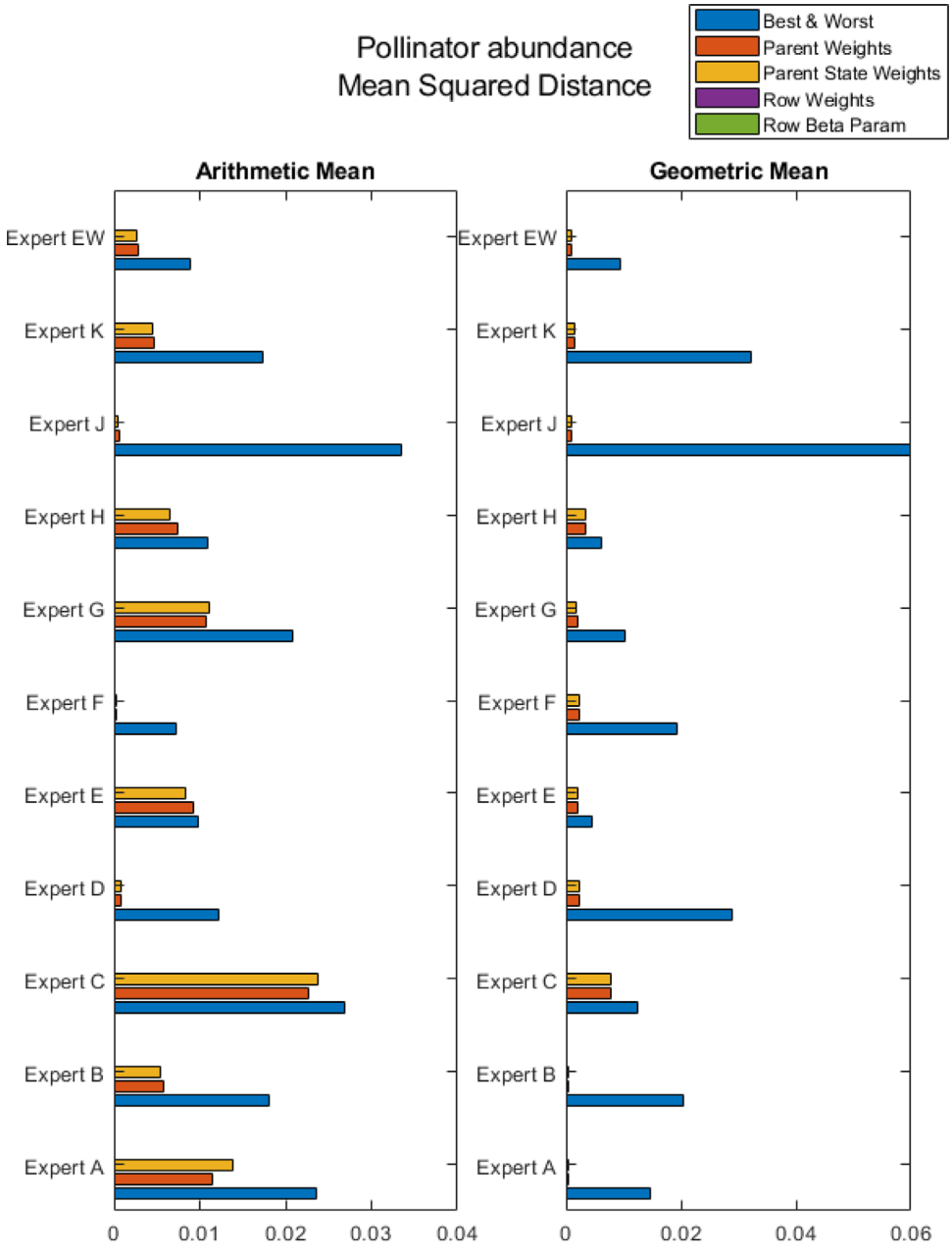

Fig A10. Bees—MSD—Arithmetic and Geometric Mean—Comparison.

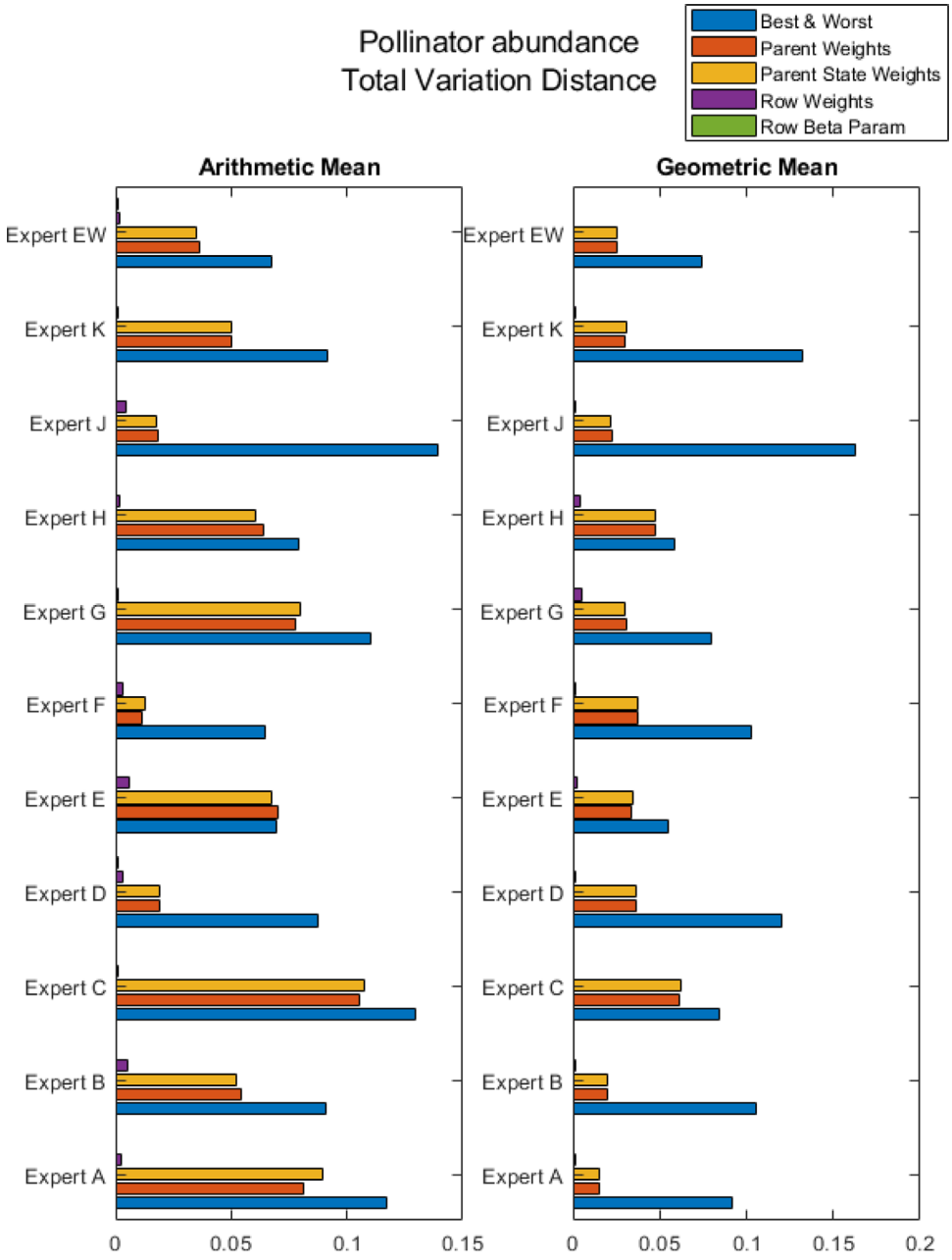

Fig A11. Bees—TVD—Arithmetic and Geometric Mean—Comparison.

REFERENCES

Alkhairy, I., & Low-Choy, S. (2017). Designing elicitation of expert knowledge into conditional probability tables in Bayesian networks: Choosing scenario. In *Proceedings - 22nd International Congress on Modelling and Simulation, MODSIM*.

Barons, M. J., Hanea, A. M., Wright, S. K., Baldock, K. C. R., Wilfert, L., Chandler, D., … Carreck, N. L. (2018). Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation. *Journal of Apicultural Research*, *57*(5), 593–604.

Barons, M. J., & Kleve, S. (2021). A structured expert judgement elicitation approach: How can it inform sound intervention decision making to support household food security?. *Public Health Nutrition*, *24*, 2050–2061.

Cain, J. (2001). *Planning improvements in natural resources management*. Wallingford, UK: Centre for Ecology and Hydrology.

Díez, F. (1993). Parameter adjustment in Bayes networks. The generalized noisy OR–gate. In Heckerman, D., & Mamdani, A. (Eds.), *Uncertainty in artificial intelligence* (pp. 99–105). San Francisco, CA: Morgan Kaufmann.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., ..Manning, B. (2016). $I_{nvestigate}D_{iscuss}E_{stimate}A_{ggregate}$ for structured expert judgement. *International Journal of Forecasting*, *33*(1), 267–279.

Hansson, F., & Sjökvist, S. (2013). Modelling expert judgement into a Bayesian Belief Network. A method for consistent and robust determination of conditional probability tables. Lund University

Hassall, K. L., Dailey, G., Zawadzka, J., Milne, A. E., Harris, J. A., Corstanje, R., & Whitmore, A. P. (2019). Facilitating the elicitation of beliefs for use in Bayesian belief modelling. *Environmental Modelling and Software*, *122*, 104539.

Laitila, P., & Virtanen, K. (2016). Improving construction of conditional probability tables for ranked nodes in Bayesian networks. *IEEE Transactions on Knowledge and Data Engineering*, *28*, 1691–1705.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*, 1st ed. Cambridge: Cambridge University Press.

Mascaro, S., & Woodberry, O. (2020). *A flexible method for parameterising ranked nodes in Bayesian networks using beta distributions*. Technical report, Bayesian Intelligence.

Nunes, J., Perkusich, M., Pereira, L., Gorgonio, K., Almeida, H., & Perkusich, A. (2018). *Issues in the probability elicitation process of expert-based Bayesian networks*. http://10.5772/intechopen.81602

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo: Morgan Kaufman Publishers.

Taguchi, G., & Konishi, S. (1987). *Taguchi methods: Orthogonal arrays and linear graphs-tools for quality, engineering*. Dearborn, MI: American Supplier Institute.

Weber, P., Medina-Oliva, G., Simon, C., & Iung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, *25*, 671–682.

Werner, C., Bedford, T., Cooke, R., Hanea, A., & Morales-Napoles, O. (2016). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research*, *258*(3), 801–819.

Whitney, C. W., Lanzanova, D., Muchiri, C., Shepherd, K. D., Rosenstock, T. S., Krawinkel, M., ..Luedeling, E. (2018). Probabilistic decision tools for determining impacts of agricultural development policy on household nutrition. *Earth's Future*, *6*(3), 359–372.

Wisse, B. W., van Gosliga, S. P., van Elst, N. P., & Barros, A. I. (2008). Relieving the elicitation burden of Bayesian belief networks. In *Proceedings of the Sixth UAI Conference on Bayesian Modeling Applications Workshop - Volume 406*, BMAW'08, p. 10–20, Aachen, DEU. CEUR-WS.org.