BSL MANAGEMENT SUPPORT
Business Simulation · Learning · Management Science

Tapping into the Wisdom of the Crowd

A Bayesian Approach to Social Forecasting

White Paper

Title: Tapping into the wisdom of the crowd: A Bayesian approach to social forecasting
Author: Guido W. Reichert
Series: White Paper No. 180223
Version: August 5, 2019

Contents

Coming to Terms with an Uncertain Future 4

Measuring Uncertainty 5

In the Eye of the Beholder 5

Give Me Five 5

Are You Certain? 7

Avoiding the Perils of Self-Deception 8

Biased Answers 8

Assessing Estimations 8

Training Estimators 9

The Wisdom of Crowds 10

Crowds or Experts? 10

Finding Common Ground 10

Pooling Estimators 11

Finally – a Glimpse into the Future 12

Selected References 13

"Our job is not to predict the future, but to prepare for it."

– Pericles

Coming to Terms with an Uncertain Future

Taking calculated risks lies at the very heart of any business. Ideally an organization's management will want to take risks in an informed and rational way. Doing so requires a look at uncertainty in its entirety – simply betting upon a single most likely prediction will not do.

These days masses of digital data help to precisely account for uncertainty and variation. Unfortunately, even when data are available, they are often not ready for use and need to be adapted to the issue at hand. Within the realm of important management issues like pricing strategies, capital investment decisions, product development, market entries and the like there are still white spots on the map where reliable data are much harder to come by.

For such issues, where data are rather scarce or not readily available at reasonable costs, we should look for good and solid methods to build upon what we already know, taking into account that what we know is rather uncertain. We would thus be well advised to see these requirements met:

- what we already know will be explicitly stated and included in the analysis,
- uncertainty pertaining to our present knowledge will be made fully transparent for risk analysis,
- precautions will be taken against biases in our judgments,
- the predictive power of groups can be exploited,
- our prior assumptions will be updated in a consistent manner as actual data become available.

Dealing with uncertainty in a business environment should start with making implicit *prior knowledge* available and putting it to good use: Your sales force, for example, is out there dealing with competition and customers daily. Why not use their valuable knowledge and judgment? Senior managers in different departments will know quite a bit about how things actually work and play out in your organization and its particular markets – expertise we would be well advised to exploit in decision making.

In this paper we will address what is called "elicitation" of *uncertain judgments* regarding quantities of interest. Most often we will be interested in inputs for a computer model that quantitatively supports decision making.[1] Growth rates for customers and sales or the number of potential buyers are typical examples for uncertain quantities of interest. An analysis of current levels and future development of intangible resources like capabilities and reputation will also benefit from expert judgment. Including intangible assets in quantitative models should quite generally enhance the value of strategic analyses.

[1] The range of models may extend from ubiquitous spreadsheets to very elaborate business simulation models.

> *"I have approximate answers and possible beliefs and different degrees of certainty about different things."*
>
> – Richard Feynman

Measuring Uncertainty

In the Eye of the Beholder

Bayesian statistics extends the narrow frequentist interpretation of probability[2], which aims to be objective and empirical, to a more general, epistemological interpretation, e.g. subjective uncertainty. After all, even in throwing dice the physical system itself is deterministic, not probabilistic; it is merely too complex and nonlinear for any observer to reliably "know" about the outcome. A lack of definite knowledge about a product's future success or the development of political climate – both clearly not identically and infinitely repeatable experiments – are cases in point for this more general interpretation of probability and uncertainty.

Following DE FINETTI [5], we may interpret a Bayesian probability as *the amount we would be willing to bet on a certain value or range of values to be true, if we were to get one monetary unit for being right, while not getting anything for being wrong.*[3] This interpretation fits business applications very well. We will use Bayesian statistics to make such bets in coherence with our prior knowledge and the available evidence, e.g. the information we have when making the bet.

Give Me Five

Let us assume, that we want our sales force to give their judgment about a new product's future market share. We want them to provide a probability distribution indicating their *degree of belief* for a range of plausible values. Since the range of plausible market shares is continuous, we would expect these distributions to be continuous as well.

While we might essentially use any continuous probability distribution as long as estimators believe them to be fair representations of their uncertainty, we specifically suggest to use a more general form of the triangular distribution, widely used in cost risk analysis and project management (e.g. [7]).

We would accordingly ask estimators to give five key values framing their judgment:

- a minimum and a maximum value defining the range of plausible values (min, max),
- a lower and an upper bound of a credible region $(\mathrm{CR}_{\mathrm{min}}, \mathrm{CR}_{\mathrm{max}})$ to contain the true value with a defined degree of certainty,
- the single most likely value (mode).

While the range for many continuous distributions like the ubiquitous normal distribution is infinite, we would like the plausible ranges to be bounded for practical reasons. The probability of the true value being outside of the range of plausible values should just be sufficiently close to zero as to be negligible. To make this more tangible think of throwing dice: The probability for any given range to be false, e.g. the true value is outside, should be lower or equal to obtaining three sixes in a row throwing a single die three times (⚅⚅⚅). In other words, the plausible range should contain the true value with at least 99.5 percent certainty.[4]

The credible region given by an estimator will be interpreted as an *interval of highest density* (HDI). Unlike an equal-tailed interval the highest density

[2] It is most often strictly defined as the limit of the frequency of occurrence for some event in an identically and infinitely repeated experiment.

[3] In using the mathematical notation $\Pr(X)$ we may thus think of *price* as well as probability.

[4] In the case of a standard normal distribution an expert may thus have given an interval around what is called the *six sigma* range, e.g. $(-3, +3)$.

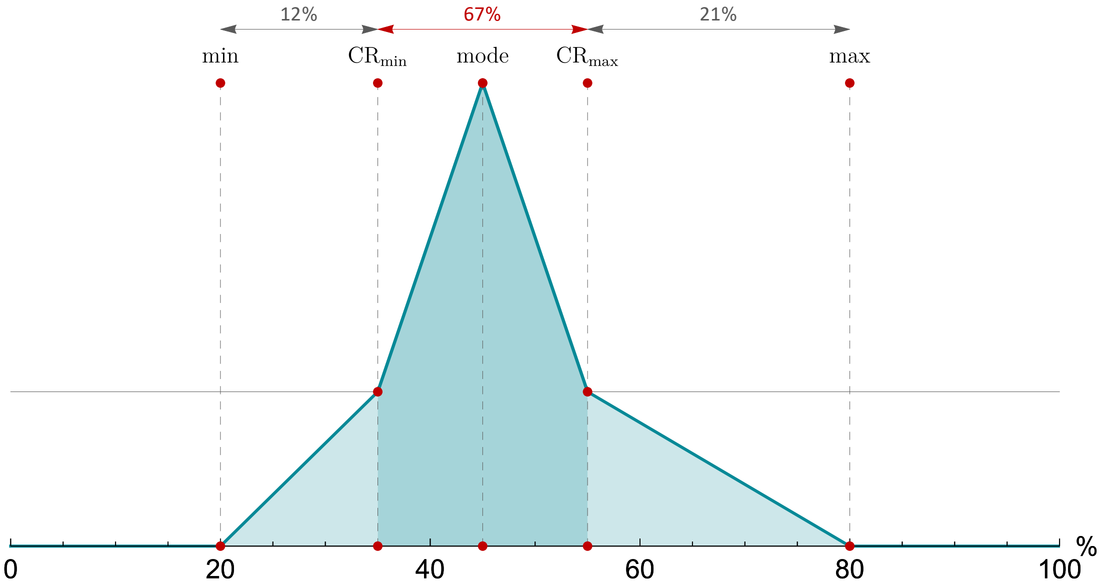


Figure 1: Estimating Market Share Using a Generalized Triangular Distribution (Pentagon Distribution)

interval is the *shortest* interval to contain the required probability. All values outside this interval will have lower probabilities – a rather reasonable property.[5]

Estimators will be given a choice of three credibility-levels for their credible-intervals, as shown in Table 1. Since they should at least be *mildly surprised* if the true value were outside their intervals, the lowest credibility-level given should surpass a mere 50:50 chance. Being twice as certain as not for the true value to be contained is an intuitively reasonable choice that immediately links to a one sigma interval for a normal distribution [13, pp. 239 ff.].

Again, providing *dice-analogies* will make assigning credibility-levels more tangible for estimators: The probability for giving a false credible-interval (e.g. one not containing the true value) should be equivalent to throwing a five or six with one die for a 67 percent interval and to throwing a ten-sided die to get ten for a 90 percent one.

Table 1: Credibility Intervals

Value contained		not contained
Probability	Chance	Dice-Analogy
67 %	2:1	⚄ or ⚅
83 %	5:1	⚅
90 %	9:1	🎲 → 10

In asking for the most likely value we implicitly require estimators to give their best guess under the premise that they would be paid a premium if the true value turns out to be *approximately equal*[6] to their estimate, but none in all other cases.

Table 2: Estimated Market Share [%]

min	CR$_{min}$	mode	CR$_{max}$	max
20	35	45	55	80

Table 2 shows the values given by a sales manager for the future market share in our example.

[5] Highest Density Intervals will therefore always include the value (or values) with the highest probability density, while equal-tailed intervals will always contain the median.

[6] We ask estimators to think of a ± 2.5 percent tolerance.

Since a 67 percent credible-interval was indicated, the distribution representing the manager's judgment would take the shape shown in Figure 1. We clearly note, that the tail probabilities are not equal and that all values outside the credible interval have lower probability densities.

Are You Certain?

It does not appear reasonable to *force* estimators into giving judgments. Depending on their knowledge about an issue and their subjective confidence estimators may feel inclined to only give a range of plausible values and no other estimate, e.g. a uniform distribution. It should similarly be admissible to simply provide the two estimation intervals but no single most likely estimate.

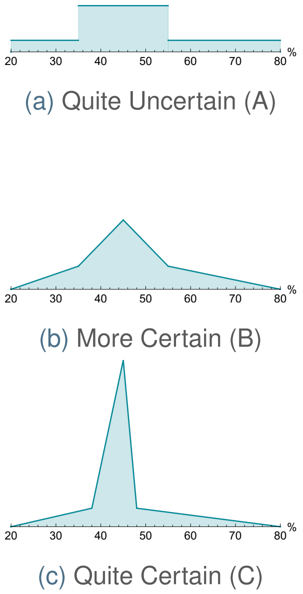
To illustrate this, suppose we had asked three estimators A, B, and C about the future market share, with their estimates given in Table 3. We recognize the values for B to be identical with those given before in Table 2. All three experts had indicated a 67 percent certainty for their credible-intervals.

Table 3: Estimated Market Share [%]

Expert	min	CR_{min}	mode	CR_{max}	max
A	20	35	–	55	80
B	20	35	45	55	80
C	20	38	45	48	80

Plotting the corresponding distributions[7] to scale, as shown in Figure 2, immediately reveals the different degrees of certainty for the experts: The more a distribution is concentrated around a single value, the more certain an estimator feels in his judgment – clearly visible in the higher probability attributed to a given most probable value.

This would be very useful information if we can trust the subjective certainties of the estimators to be justified. But can we?



(a) Quite Uncertain (A)

(b) More Certain (B)

(c) Quite Certain (C)

Figure 2: Estimates for Market Share [%]

[7] If no mode is given the distribution will simply be a mixture of two uniform distributions.

"What I would eliminate if I had a magic wand? Overconfidence."

– Daniel Kahneman

Avoiding the Perils of Self-Deception

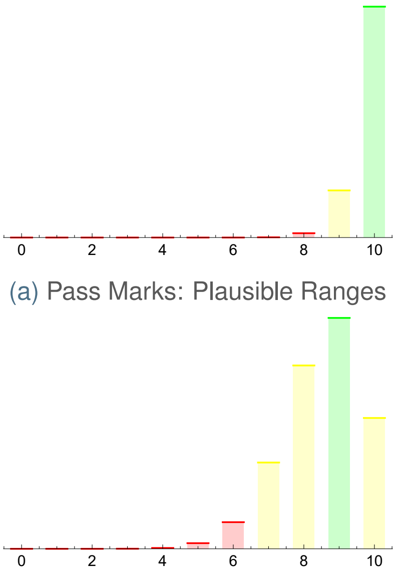
Biased Answers

We had described the link between Bayesian probability and betting odds in the last section. Let us assume that we were to pay a premium to experts whenever a true value correctly falls within their intervals. In the case of a 90 percent credible interval experts should ideally be completely indifferent between throwing a ten-sided die once to get the premium in case a number in the range of one to nine appears and betting on the correctness of their interval.

Such considerations have given rise to various exercises and tests of an estimator being *well calibrated* [11]. Most often almanac questions are used (e.g. "How many DVDs are currently sold in the EU per year?") to test whether an estimator is calibrated. Exposing estimators to calibration exercises almost always reveals a great amount of *overconfidence*, e.g. out of ten test questions on average only six or seven plausible ranges and only four or five 90 percent credible intervals will contain the true value. The culprit quite often is *anchoring*: estimators stick to readily available but false initial point estimates. They thus fail to systematically decompose problems and to start out from ranges and credible intervals. [15, pp. 100 ff.]

Assessing Estimations

The assessment of plausible ranges is straightforward: a calibrated estimator's range of possible values should contain the true value at least with a probability of 98 percent, giving some allowance. Similarly, highest density intervals of 90 percent, which are either explicitly stated or implicitly defined by the given distributions, should contain the true value with probabilities between 85 and 95 percent to be in accordance with calibration.



(a) Pass Marks: Plausible Ranges

(b) Pass Marks: Credible Intervals

Figure 3: Assessment for Ten Test Questions

If ten test questions are given, then Figure 3 will show probabilities for the number of correct intervals assuming the true probability is such that it can be accepted just so, e.g. 98 percent for plausible ranges and 85 percent for credible intervals. In the diagrams red columns indicate a probability of less than five percent, green ones expected results. As we can see, not achieving more than six correct credible intervals out of ten test questions would not be in accordance with calibration (Figure 3b).[8]

The assessment of point estimates is less obvious. At this stage we are not concerned with the knowledgeability of estimators, instead we care about their ability to realistically express their beliefs using probability distributions. The question we should thus be asking is:

[8] There are, of course, interdependencies, so in Figure 3b we will want to examine only those test questions where the plausible ranges were given correctly.

> “Has the point estimate turned out to be *useful* or would we have been better off without it?”

Figure 4 shows the credible interval for the market share estimate as given in Table 2. The horizontal orange line at medium height marks the level of credibility if no point estimate had been given. Thus, if in our example the true market share turns out to be between 40 and 50 percent then the estimator would have – justly – allocated more credibility upon these values by giving a point estimate (green area). Conversely, a true value outside of this range indicates that it would have been better not to have given a point estimate (red area).

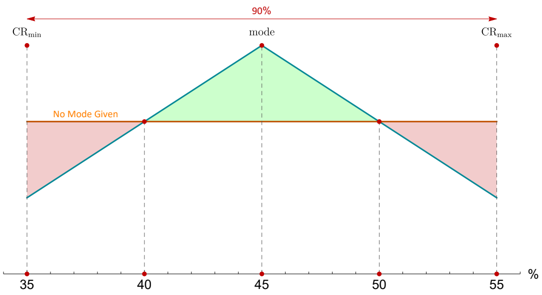


Figure 4: To Give or Not to Give?

Assuming that a given credible interval is correct, the odds are 2 to 1 against providing a useful point estimate by just randomly picking a value within the interval. From this fact we can develop an assessment similar to those above. Figure 5 shows that when seven credible intervals were correctly given – which is what we would expect for 67 percent intervals – we would find it very unlikely that an estimator with six or more useful point estimates had been guessing at random. As it turns out, the pass mark of at least six useful point estimates does also hold for more than seven correct credible intervals.

Training Estimators

It has been shown that the ability to provide correct estimation intervals will improve with training if the estimators are given feedback [9, pp. 111-118]. For this task we have developed a software called Panel Estimation Tool (PESTO), which is cloud-based. It will record estimations and in the calibration phase immediately show, which estimates were wrong – providing the true answer as well. The tool will also report the correct/useful estimates as moving averages to indicate the progress being made. A sufficient level of calibration will in most cases be reached within 30 to 40 questions. Spending some time on calibration training therefore is a worthwhile investment.

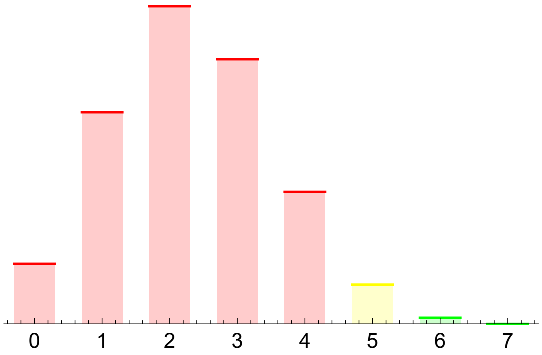


Figure 5: Assessing Useful Point Estimates

One should think of almanac questions and so called “guesstimates” as a kind of worst case scenario for testing as domain experts will have no advantage to laypersons.Guesstimates test the ability to apply “street-fighting mathematics” [12] and good reasoning – a transferable skill with regard to business issues. We can be reassured that experts will show better performances within their domains of expertise [15, p. 73].

> *"Diversity and independence are important because the best collective decisions are the product of disagreement and contest, not consensus or compromise."*
>
> – James Surowiecki

The Wisdom of Crowds

Crowds or Experts?

GALTON [6] famously discovered that averaging over a pool of independently given estimations turns out to provide a superior estimate when he visited a local livestock fair in 1906. The "wisdom of crowds" has been confirmed many times over ever since (e.g. [16]).

The surprising message of these findings is, that the quality of a group's estimate does not depend upon the amount of experts within it. On the contrary, the more diverse the group the better the estimate. Group interaction, e.g. the Delphi method and other techniques, does not seem to provide any advantages to independent elicitation of judgments and applying a consistent mathematical procedure for aggregation [15, pp. 190 f.].

In another line of research started by PAUL MEEHL [14] experts intuitively grounding their judgment on huge amounts of information lose out to rather simple algorithms using muss less information – albeit in a consistent fashion. This result has been confirmed rather impressively by a large meta study [8].

What are we to make of these results? At the least, we should be skeptical of – sometimes very elaborate – procedures of ranking estimator-performances to find calibrated *and* knowledgeable experts (e.g. Cooke's Classical Method [3]). The questions, When should we stop the search? or Why ask more than one, top-ranking expert? are not answered to our satisfaction.

In business settings it should suffice to identify between 10 and 15 estimators who are seen as competent by their peers or the decision maker for any unknown quantity or a cluster of similar quantities of interest. From the above findings we should conclude that we might be better off to err on the inclusive side of the problem – in any case, we should make sure of sufficient diversity in our group of estimators so that different sources of information are included. Such a group of experts will be referred to as the *expert panel*.

Finding Common Ground

Bayesian probabilities are always conditional; in the case of prior estimates, subjective probabilities will be dependent on whatever information is available to estimators: prior market research, personal experience, or knowledge. Probability judgments will also dependent upon the context they are made in, e.g. the specific model used to support decision making and its informational needs.

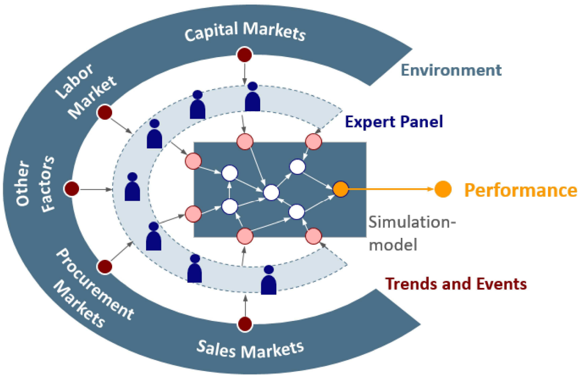


Figure 6: Background Information

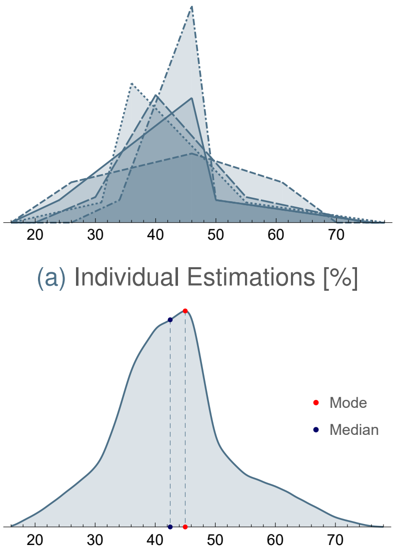
We would like to establish some kind of common ground – a core of shared background information – for all estimators participating in an elicitation, while heeding the call for independence and diversity. Typically, we would first present the chosen model to support decision making, that is if it has not been developed in cooperation with the expert panel already. We will be prepared to adapt a given model if discussion following the presentation produces good reasons to do so. Having reached acceptance and understanding with regard to the model and its informational needs, we will use questionnaires, interviews, and an additional group discussion, if needed, to scan whatever trends and events experts note in the environment with relevance for model input. Figure 6 illustrates this for broad, strategic issues – other types of decisions will have a more narrowly defined background.

Available background information will then be processed and summarized to provide some kind of reminder in written form, e.g. a "crib sheet". During the elicitation phase experts can then let the summarized background information "resonate" with their gut feelings. For elicitation the same cloud-based software (PESTO) is used as for calibration. Experts can either work on their own – and at their leisure – or give their estimates within a structured interview. While the first option is very flexible, the last one can give better results as interviewers will challenge estimations and thereby ensure sufficient reflection.

Pooling Estimators

Once all estimations are given, the individual distributions will have to be combined to provide panel estimations. A widely accepted and robust procedure for doing this is a weighted, linear combination of all distributions for a quantity of interest [2].

But, which weight should be given to individual estimations? This question is a hot one and far from having a clear answer. Cooke and Goossens [4], to give a prominent example, have devised a quite elaborate procedure to weigh estimators. Unfortunately, the theory behind this is not very strong: How to reliably test an expert about things we do not know about – especially in a business setting? Weighing experts risks introducing a false bias and may not necessarily prove to be worth the effort (see [1]). We therefore suggest to give all experts equal weights, following the classical example of Galton above and widely accepted practice [15, p. 191].



(a) Individual Estimations [%]

(b) Panel Estimation [%]

Figure 7: Combining Judgments for Market Share

Figure 7a shows five estimations obtained from experts for future market share. The equally-weighted linear average of these estimations will result in the panel estimation shown in Figure 7b. Its most probable value (mode) turns out to be 45 percent, while its median value turns out to be around 42 percent. The panel estimation's range will be the union of all the plausible ranges given and thus be maximal. The probability distributions obtained for the group will account for all the information provided by the experts.

"It is better to be vaguely right than exactly wrong."

– Carveth Read

Finally – a Glimpse into the Future

By now we have managed to independently elicit uncertain judgments from calibrated domain experts and to combine this information to form panel estimations for all quantities of interest. We can now use this prior knowledge in our model to quantitatively support decision making.

Figure 8 illustrates the use of expert judgments for predicting an organization's future performance. As far as only prior information (expert judgments, data) is used, available at the time of making the prediction, the result will be a *prior predictive distribution*. Data that eventually become available can later be used to *update* our prior distributions to obtain a *posterior prediction*. In such a case, Bayes' rule will guarantee that new information is accounted for consistently.

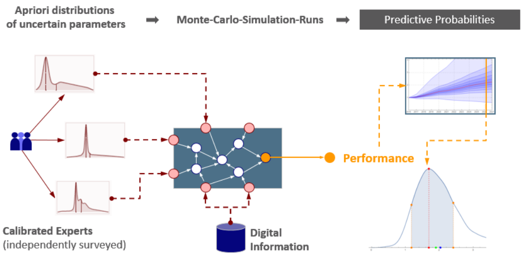

Apriori distributions of uncertain parameters
Monte-Carlo-Simulation-Runs
Predictive Probabilities
Performance
Calibrated Experts (independently surveyed)
Digital Information

Figure 8: Prior Prediction of Future Performance

Our methodology therefore meets all of the requirements that we had listed in the beginning. Since we will be working with full probability distributions from start to end, we will have a smooth transition to decision theory. This will already become apparent in having a *choice* for selecting a single point value summarizing our prediction for a specific point in time. As indicated in the distribution function in the lower right corner in Figure 7b, we may pick the mode, the median, or the mean of the predictive distribution for a specific point in time; the best choice will then depend upon the implication a likely prediction error will have for us [10, pp. 63-70].

But we can go further: Using elicitation *conditional probabilities* can be obtained from experts as well. This will be relevant whenever we want to model uncertain outcomes of actions and events in decision models. An immediate question to ponder might be whether it is reasonable to invest in more information, for example market research, and how much we should be willing to pay for additional data (see [9]).

Expert judgments can also be used to elicit *multivariate distributions* so that we may address dependent quantities of interest. This will be relevant if we want to estimate proportions adding up to 100 percent, e.g. market shares for multiple competitors, or if we use marginal distributions of income, age, and other attributes to estimate the joint distribution for a demographic population of interest, e.g. our customers or clients.

Using the knowledge available within our organization in this way should become as natural as opening up a spread sheet. Elicitation and usage of uncertain expert judgments offers a reasonable point of entry into modern risk management, decision support, and predictive analytics in general.

Selected References

[1] Robert T. Clemen. Comment on Cooke's classical method. *Reliability Engineering and System Safety*, 93(5):760–765, 2008.

[2] Robert T. Clemen and Robert L. Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203, 1999.

[3] Roger M. Cooke. *Experts in uncertainty: Opinion and subjective probability in science*. Environmental ethics and science policy series. Oxford University Press, New York, 1991.

[4] Roger M. Cooke and Louis H. J. Goossens. TU Delft expert judgment data base. *Reliability Engineering and System Safety*, 93(5):657–674, 2008.

[5] Bruno De Finetti. *Theory of probability: A critical introductory treatment (Translated by Antonio Machì and Adrian Smith)*, volume 6 of *Wiley series in probability and statistics*. John Wiley & Sons, 2017.

[6] Francis Galton. Vox populi. *Nature*, 75(7):450–451, 1907.

[7] Lionel A. Galway. *Subjective probability distribution elicitation in cost risk analysis: a review*. RAND Corporation, Santa Monica, CA, 2007.

[8] William M. Grove, David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):19–30, 2000.

[9] Douglas W. Hubbard. *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons, Hoboken, 3rd edition, 2014.

[10] Karl-Rudolf Koch. *Introduction to Bayesian statistics*. Springer, Berlin, 2nd edition, 2007.

[11] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips. Calibration of probabilities: the state of the art to 1980. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgment under uncertainty*, pages 306–334. Cambridge University Press, Cambridge, 1982.

[12] Sanjoy Mahajan. *Street-fighting mathematics: The art of educated guessing and opportunistic problem solving*. MIT Press, Cambridge, Mass. and London, 2010.

[13] Sanjoy Mahajan. *The art of insight in science and engineering: Mastering complexity*. MIT Press, Cambridge, Mass. and London, 2014.

[14] Paul E. Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, Minneapolis, MN, 1954.

[15] Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, Richard J. Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons Ltd., Chichester, 2006.

[16] James Surowiecki. *The wisdom of crowds.* Anchor Books, New York, 1st edition, 2005.

BSL MANAGEMENT SUPPORT

Business Simulation · Learning · Management Science

Dipl.-Kfm.
Guido Wolf Reichert

Schauenburgerstr. 116
24118 Kiel
Germany

☎ +49 431 5606 855
🖨 +49 431 5606 856
@ gwr@bsl-support.de
🌐 www.bsl-support.de