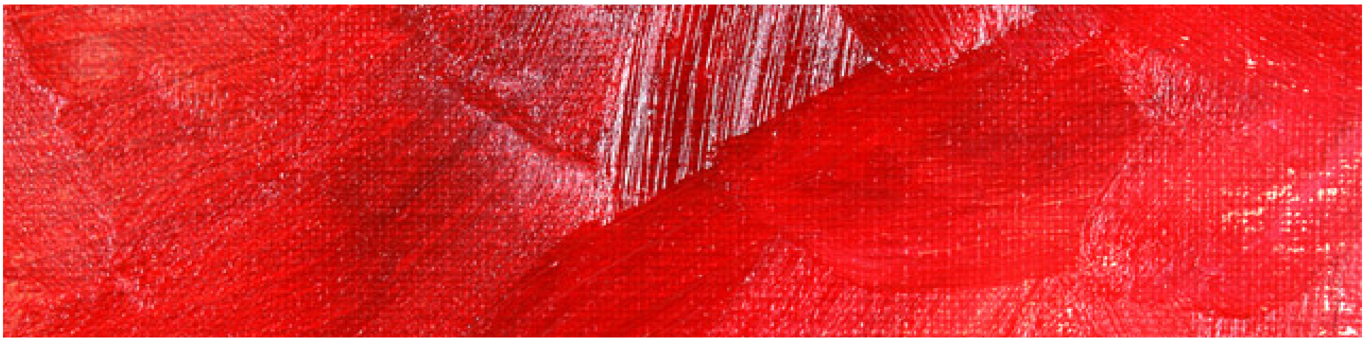
cecan
Centre for the Evaluation of Complexity Across the Nexus

Note No. 2 .1
Winter 2016

Testing Contribution Claims with Bayesian Updating

A CECAN Evaluation and Policy Practice Note for policy analysts and evaluators



Bayesian updating can be a powerful means of testing claims about policy impact. It can be a rigorous, "disciplined" addition or complement to other theory-based approaches aiming to test theories and mechanisms, and a useful tool for policy evaluators and analysts.

What is Bayesian updating?

Bayesian updating is used to formulate and test contribution claims. A contribution claim:

- Is a statement about the contribution an intervention made to an outcome, addressing how and / or to what extent the intervention contributed to it.
- Often includes the role of other factors not directly related to the intervention.
- May be tested, using Bayesian updating, to estimate the degree of confidence that should be placed upon it.
- Is a "statement about reality" (an ontological statement). However, the idea that the claim is true or not is a hypothesis "in our head" (Bennett & Checkel, 2014)

How does Bayesian updating work?

The researcher/evaluator

- Initially holds a degree of confidence about the claim/statement being true or not, on the basis of logic, theory or prior knowledge: this is known as "the prior".
- Updates the confidence in the claim on the basis of empirical observations or new emerging evidence. In order to do this the researcher/evaluator needs three items of information:
    - The probability of the claim being true before observation of the new evidence ("the prior").
    - The probability of observing the evidence if the claim is true ("the sensitivity").
    - The probability of observing the evidence if the claim is false ("the type 1 error").

Where/when can Bayesian updating be used most effectively to test contribution claims?

Bayesian updating can usefully be applied:

- In a number of different areas and sectors, including medical diagnosis, law, crime investigation, forensic science, historical studies, political science, geology and archaeology.
- When contribution claims include components and mechanisms which cannot be directly observed and measured, making them difficult to test, particularly in complex settings.
- When qualitative claims and the existence of mechanisms need to be tested.
- When a high degree of transparency and internal validity is required. The probability estimates are subject to open and transparent scrutiny: anyone can challenge the proposed values and argue for alternatives.
- When it is necessary to build stakeholder consensus and to enhance the credibility of contribution claims: if experts or stakeholders end up agreeing on a set of values or intervals, the agreed confidence in the claim can be considered robust.
- In evaluation, as a rigorous formalisation of "process tracing" (Befani et al., 2016; Befani & Stedman-Bryce, 2017).

What preparation is needed in order to apply Bayesian updating?

When planning the use of the method it is important to bear in mind that:

- At least one member of the evaluation team needs to have (or gain) an understanding of, and ability to apply, Bayesian logic.
- Because the credibility of the method rests on the reliability of the probability estimates it is of paramount importance that a credible method for the estimation of probabilities is used.
- Precise point estimates are not necessarily needed: in some cases, the posterior probability's interval is relatively small even when other probabilities' intervals are large.
- Probabilities can be estimated on the basis of: 1) existing empirical data; 2) computer-based simulations; 3) expert input.

What problems may arise in using Bayesian updating?

In many real-life evaluations there is insufficient information to calculate probabilities directly, so we rely on the judgement of experts to make a subjective assessment (see Cook, 2001; Gosling, 2014; Hora, 2007; O'Hagan, 2010). However expert judgements can be biased because:

- The expert's response is/might be motivated by personal interests or "motivational biases". They may be inclined to over-emphasise the probability that a contribution claim is true if they have a personal stake in its acceptance.
- They may feel that because they have been identified as an "expert" they should provide a clear opinion (e.g. "the probability is 90%") whereas in fact a greater degree of uncertainty is more realistic.
- When the expert processes their underlying knowledge they may think in certain ways that introduce a systematic adjustment or "cognitive bias". Some common cognitive biases that particularly affect subjective probability assessment include:
  - Anchoring – focussing on a specific "official" number.
  - Availability – focussing on a recent or particularly memorable case.
  - Coherence – putting excessive weight on a "good story" at the expense of more complex or ambiguous explanations.
  - Overconfidence – underestimating the probability of an unusual event.

What approaches can overcome biases and ambiguity?

A structured approach to expert judgement seeks to minimize any biases and sources of ambiguity, and ensures that the process is as transparent as possible. Such an approach will:

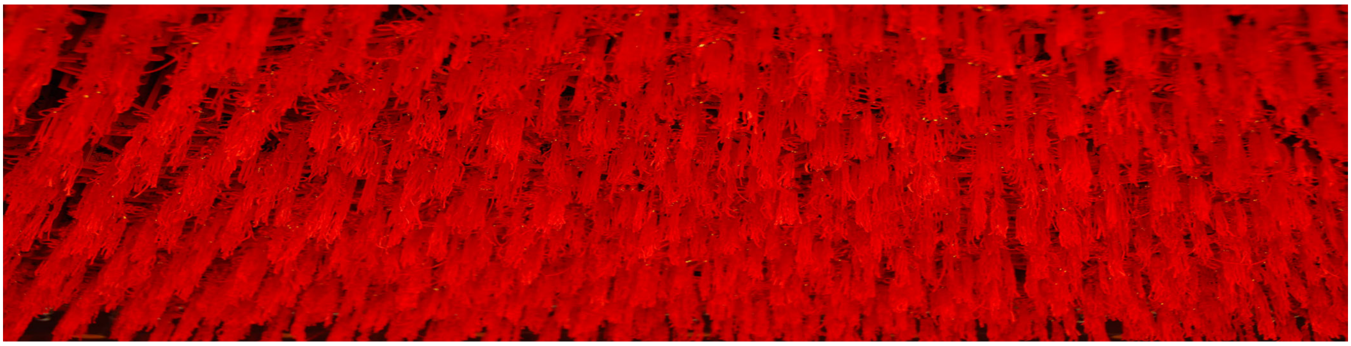
- Separate the role of "expert" from the role of "facilitator", whose job is to ensure that the process is well run.
- Allow the facilitator to use various tools to help the expert decide on a probability value, for example comparing the current situation to lotteries with known probabilities such as coin tosses. However, perhaps the most useful technique is to use a probability scale such as that used by the IPCC for assessing climate change evidence:

Term	Likelihood of the outcome
Virtually certain	99-100%
Very likely	90-100%
Likely	66-100%
About as likely as not	33-66%
Unlikely	0-33%
Very unlikely	0-10%
Exceptionally unlikely	0-1%

Source: Technical Summary In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change

- May use a panel of experts rather than rely on a single opinion, but that introduces its own difficulties. Even when experts start from a common understanding of the question to be addressed and have access to the same sources of information they may still have differing opinions of probability. Therefore, most approaches to subjective probability assessment try to generate a single number that represents a synthesis of the group's opinions.

- Can combine individual expert opinions into a group view. Two common approaches are used:
  - Consensus based approaches that argue a group will eventually come to a consensus opinion about the correct probability or range of probabilities to use through discussion.
  - Mathematical pooling approaches that argue individual expert opinions should be combined mathematically using weighting factors. The simplest approach is to assume equal weights for all experts, but performance based weights are also used, for example with weights calculated based on the expert's answers to calibration questions that have known outcomes.



Case study: impact evaluation of a health care advocacy campaign in Ghana

- The refined contribution claim that was tested with Bayesian probability is “the report proposed a formula for estimation of health system coverage which was eventually adopted by the Government”.
- Various pieces of evidence with high sensitivity were sought in order to decrease confidence in the claim (i.e. to show that the campaign had likely not influenced policy), none of which were found.
- Finally, two pieces of evidence with low type I error were used to increase confidence in the claim: the perfect match between the formula proposed and the formula adopted and the admission of influence on behalf of the Government of Ghana in a tense and conflictual context where such an admission was not expected. For the latter piece of evidence, three scenarios were formulated and – at least in that specific context – it was considered much more likely for the Government to admit influence in case of influence, than to admit influence in case of no influence. The initial probability of 0.5 was updated to 0.95 in the standard and 0.75 in the conservative scenarios.
- The main weakness of this study was that – although a useful and clarifying exercise – quantifying probabilities was largely subjective and should have drawn on more established procedures of subjective probability estimation.

Case study: Impact evaluation of an information exchange network on poverty and conservation

- Two hypotheses were tested:
  - That the network had contributed to shape the content of a policy decision by the Uganda Wildlife Authority (about increasing the community share of the revenue from the gorilla permit fee);
  - That the same network had accelerated the process.

Two mechanisms were formulated describing the influence processes and the roles of the network.

- Several pieces of evidence were considered and assessed for each mechanism component, with independent pieces of evidence bulked together into packages, and leading to confidence updating for the existence of each mechanism component.
- The mechanism components with the highest confidence were those for which meeting minutes and email transcripts were available, as opposed to accounts collected during interviews.
- An overall likelihood score was assigned to each mechanism on the basis of the lowest confidence score of its component: in other words, confidence in the overall mechanism was only as good as confidence in its weakest link.
- The main weakness of the study was: estimation of subjective probabilities could have potentially been conducted in a more structured and transparent way, while the timing of the evaluation did not afford the team with such opportunity.

Case study: Energy policy evaluation analysing the decision making behaviours of consumers regarding the uptake of the FITS energy scheme

- The scheme sought to incentivize the installation of domestic photo voltaic panels for local hot water and electricity production. There was an announcement 6 months in advance that a subsidy for which users were applying was to be removed, and applications saw a sharp drop after the announcement.

- The evaluation sought to test the claim that the drop was due to the dissemination of news about the impeding subsidy removal; in particular that the news had triggered a negative reaction to the news (which was simulated) as opposed to a no reaction or a positive reaction (which would have seen a last minute raise in applications).

- An agent based model was developed and validated with DECC figures for actual subsidies issued, which the model attempted to replicate with a simulated society. Bayesian Updating was not considered for this work and so none of the probabilities were estimated; if it had been considered, in addition to the three priors, the pieces of evidence for the claims could have been the difference between the number of applications filed during the three months after the news compared to the same time period in the previous year; and the number of applications filed during the three months before the news compared to the same time period the previous year. Running the model repeatedly under each of the three hypotheses (no reaction, positive reaction, negative reaction) would have returned average values of the number of applications during those time periods under the different hypotheses. This would have allowed the estimation of sensitivity and type I error for each piece of evidence under each claim, and eventually updated the priors into the posterior probabilities of each hypothesis, where the expectation would have been that the most strongly supported hypothesis, the one with the highest posterior, would be the "negative reaction to the news".

How can Bayesian Updating be useful in evaluation and how might it be developed further?

- Given its application to a high number of fields, Bayesian Updating could be a useful tool for evaluation. Attempts to use it so far have shown that:
- The method can be a powerful opportunity when unobservable and qualitative contribution claims – which are often encountered in the evaluation of complex programmes in uncertain settings – need to be tested.
- It can be a rigorous, "disciplined" addition or complement to other theory-based approaches aiming to test theories and mechanisms, like contribution analysis and realist evaluation; and is – for all practical purposes – a quantitative formalisation of Process Tracing.
- The main critical issue rests in the estimation of probabilities, for which a set of three possibilities are:

1) use empirical data if available;

2) use computer-based simulations if possible;

3) estimate subjective probabilities using either consensus based or mathematical pooling approaches.

- Further development might include guidance on how to estimate the above probabilities, particularly when pulling together evidence from multiple sources; and when the sources are both empirical and subjective. In addition, much needs to be learned concerning how/under what circumstances sources can be considered stochastically independent. More generally, practical and user-friendly spreadsheet tools incorporating all the necessary information and estimates need to be developed in a format which is ready for use by evaluators and practitioners.

References and further information

Befani, B., D'Errico, S., Booker, F. & Giuliani, A. (2016) "Clearing the fog: new tools for improving the credibility of impact claims", London: International Institute for Environment and Development. Available at http://pubs.iied.org/17359IIED/

Befani, B. & Stedman-Bryce, G. (2017) "Process Tracing and Bayesian updating for impact evaluation", forthcoming in Evaluation; available OnlineFirst at http://evi.sagepub.com/content/early/2016/06/24/1356389016654584.abstract

Bennett, A. & Checkel, J. (2014) "Introduction: Process tracing: from philosophical roots to best practices", in: Bennett & Checkel (eds) Process Tracing: From Metaphor to Analytic Tool Cambridge University Press.

Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change

Cooke R.M. (2001), Experts in Uncertainty, Oxford University Press;

Gosling J.P. (2014), Methods for eliciting expert opinion to inform health technology assessment;

Hora S.C. (2007), "Eliciting probabilities from experts", in Advances in Decision Analysis: From Foundations to Applications, Cambridge University Press;

O'Hagan A. (2010), SHELF: The Sheffield Elicitation Framework, available on-line at http://www.tonyohagan.co.uk/shelf/index.html

cecan
Centre for the Evaluation of Complexity Across the Nexus

www.cecan.ac.uk / cecan@surrey.ac.uk / +44 (0) 1483 682769

The Centre for the Evaluation of Complexity Across the Nexus (CECAN) is a £3m national research centre hosted by the University of Surrey, bringing together experts to address some of the greatest issues in policy making and evaluation.

This Evaluation Policy and Practice Note was written by Barbara Befani, Chris Rees, Liz Varga and Dione Hills
Contact: befani@gmail.com / b.befani@surrey.ac.uk

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
Department for Environment Food & Rural Affairs
Department for Business, Energy & Industrial Strategy
NERC SCIENCE OF THE ENVIRONMENT
Environment Agency
Food Standards Agency food.gov.uk