Defining and identifying uncertainty

Defining uncertainty

Sources of analytical uncertainty - Data

Sources of analytical uncertainty - Assumptions

Sources of analytical uncertainty - Analysis

3.

3. Defining and identifying uncertainty

Analytical uncertainty can feed through into analysis and subsequent decision making from many different sources

We encounter uncertainty throughout the decision making process and in the analysis which supports it. In addition to uncertainties around the analytical question, we will also find uncertainty in the context of the decision being made, the data and assumptions feeding into the analysis and in the analysis itself. As analysts we need to understand and describe contextual uncertainties to ensure our analysis has impact; and we need to describe and quantify analytical uncertainties to ensure decision makers are clear about how far analytical results can be used to support their conclusions.

Early identification is important

Try to identify and record all the potential sources of uncertainty in your analysis at an early stage. Early identification of uncertainty is important; if you overlook a potential source of uncertainty this could reduce the usefulness and impact of your subsequent analysis. See the Presenting and communicating uncertainty (https://analystsuncertaintytoolkit.github.io/UncertaintyWeb/chapter_4.html section for Tornado diagrams that are a useful way to communicate the size of uncertainty.

This section sets out a range of techniques to help you understand and assess the sources of uncertainty in your analysis.

3.1. Defining uncertainty

Understanding the characteristics of different classifications of uncertainty can help you to identify sources of uncertainty in your own analysis. Further, categorising the types of uncertainty provides a framework for the next steps of analysis.

A common classification divides uncertainty into known knowns, known unknowns, and unknown unknowns. We explain these in Table 3.1.

Table 3.1: Classifications of Uncertainty

Classification	Aleatory uncertainty	Known unknowns - Epistemic uncertainty	Unknown unknowns - Ontological uncertainty

Classification	Aleatory uncertainty	Known unknowns - Epistemic uncertainty	Unknown unknowns - Ontological uncertainty
Definition	Sometimes referred to as "known knowns", aleatory uncertainty is the **things we know that we know**. This refers to the inherent uncertainty that is always present due to underlying probabilistic variability.	Known unknowns are **things that we know we don't know**. This type of uncertainty comes from a lack of knowledge about the (complex) system we are trying to model. Assumptions are used to plug these gaps in the absence of information.	Unknown unknowns are **things that we don't know we don't know** . It usually comes from factors or situations that we have not previously experienced and therefore cannot consider because we simply don't know where to look in the first instance.
Can it be quantified?	**Yes** it can be quantified. We usually characterise it using a probability distribution function (PDF). A PDF gives all the possible values that a variable can have and assigns a probability of occurrence to each. As analysts, the challenge for us is to derive the PDF. If you find that you can't then you may instead have a known unknown.	**Yes** it can be quantified (but isn't always) – e.g. through sensitivity analysis. These techniques try to quantify the uncertainty by altering assumptions and observing the impact on modelling outputs. They will work if the range of assumptions tested covers the range of unknown variables.	**No** it cannot be quantified. We cannot identify unknowable unknowns, so there are no actions we can take to quantify them. What we can do is be clear about the sources of uncertainty we have included, so that any others subsequently identified would likely add to that uncertainty.
Can it be reduced?	This type of uncertainty cannot be completely removed. We can sometimes reduce it through data smoothing or increasing the size of a sample, but there will always be some random variability.	Known unknowns are reducible by gathering information to lessen the gaps in our knowledge. Using new data sources, expanding our data collection or conducting research can remove the need for assumptions or refine their ranges.	This type of uncertainty is not reducible. However, this type of uncertainty can usually be separated into "unknowable unknowns" and "knowable unknowns". Horizon scanning can help identify knowable unknowns. Once they are identified they become known unknowns.
Example	Tossing a coin is an example of aleatory uncertainty. We can observe the possible outcomes (heads or tails) and the probability of each occurring (50:50), therefore create the PDF. However, prior to the coin being tossed we cannot reduce the uncertainty in outcome.	Taking our coin toss example, we don't know whether the coin is fair in the first instance. We may assume the coin is fair and will give a 50% probability of each outcome. Once we start to toss the coin, we start to gather information on its fairness. The longer we toss the coin the better our information gets and the greater the reduction in the known unknown.	Unknown unknowns are often future events or circumstances that we cannot predict, for example, somebody swaps the coin to a weighted one without our knowing, or steals the coin altogether! Previous analysis is no longer reliable as it didn't account for this change.

3.2. Sources of analytical uncertainty - Data

The data that feeds into your analysis project will have been previously specified, defined, and collected. In some cases, you will do this yourself, but you may also draw on data sources collected by others. Having chosen your data sources for your project you will need to think about how well your data describes the reality of the situation you are modelling or analysing.

To gain a full picture of the impact of data uncertainty on your analysis you should think through what you know about where your data has come from. You should use a data log with quality and impact Red Amber Green (RAG) ratings. Consider the following questions:

How your data source compares with your analysis objective:	How well do the definitions and concepts in the data chosen fit with what you are trying to measure? Differences between the data and your target group can mean that a dataset captured for one purpose is inappropriate for another. For example, you might want to analyse London & South East but only have data for the whole of the UK.
Where the data come from and how they have been collected:	How rigorous was the data collection process? Was the data owner's quality assurance sufficiently robust? For survey data, would respondents have fully understood the question intent? Some datasets are subject to regulation and compliance with standards or other codes of practice. In such cases, quality should be well documented and assured like in National Statistics.

When considering uncertainty in input data, you should think about whether the data being used was gathered for an alternative purpose and if it has been manipulated and how you can adjust or account for this. Accompanying data descriptions (or a quick exploration of the source data if these don't exist) can be helpful in understanding data limitations of the data and whether any adjustments made could conflict with or bias your analysis. Statistical sources often come with supporting information about accuracy and reliability. You can sometimes find information on variance (or standard errors, confidence intervals, coefficients of variation) and you may find indications of likely bias, from special studies comparing or linking sources. These direct measures of quality, together with indirect measures such as response and coverage rates can tell you a lot about the uncertaint.

What period the data covers:	More uncertainty will occur if either the data don't match the time period of interest and/or if the data are volatile.
Whether your data has been subjected to any pre-processing:	For data obtained in a processed state from others you may need to explore what processing steps were taken to determine how that may affect the data you are using. For example, missing values may have been imputed, survey data may have been weighted to make survey results representative of a wider population, extreme values and outliers may have been removed, data sets may have been combined (possibly resulting in false positive or false negative matches), disclosure controls may have been applied (potentially biasing the data set). Consider how the retention or exclusion of an outlier will affect your results. Truncation or removal of outliers will typically introduce bias but this may be tolerated in exchange for reduced variance.
Check whether there is any bias or uncertainty in the data	Statistical sources often come with supporting information about accuracy and reliability. You can sometimes find information on variance (or standard errors, confidence intervals, coefficients of variation) and you may find indications of likely bias, from special studies comparing or linking sources. These direct measures of quality, together with indirect measures such as response and coverage rates can tell you a lot about the uncertainty. In the absence of direct measure of variance, be aware that small sample sizes will increase the margin of error in your results.

3.3. Sources of analytical uncertainty - Assumptions

Considering the assumptions you're making in your analysis is critical to any uncertainty analysis

Consider where you have used assumptions	Assumptions are used when we have incomplete knowledge. All models will require some assumptions, so you need to ensure that assumptions are robust and consistently understood. You should use an assumptions log with quality and impact RAG ratings and they should be signed off by stakeholders. Where did the assumptions come from? How were they generated and why? What is the impact if they are wrong, and how often are they reviewed?

What assumptions are outside the scope of the model?	There are often parameters outside of the scope of the model that have been implicitly assumed. For example, models may assume no substantial policy changes in related areas and there may be deliberate limits in the coverage or timelines of your analysis – deliberate modelling exclusions that allow timely and effective analysis. These assumptions and limitations provide the context in which the modelling results are appropriate. You need to be aware of the restrictions that these assumptions impose on the interpretation of analytical results and take care to explain where modelling results can (and cannot) be used.
Assess the quality of each assumption	Assumptions should be based on robust evidence. The less evidence to support an assumption the more uncertain it will be. High quality assumptions will be underpinned by robust data, while low quality assumptions may simply be an opinion or may be supported by a poor data source.
Assess the impact of each assumption	The importance of an assumption is measured by its effect on the on the analytical output. The higher the impact of an assumption the more uncertain results will be. Critical assumptions will drastically affect the results, while less importance assumptions may only have a marginal effect on results. More weight should be given to gathering evidence to improve the quality of critical assumptions.
What don't you know?	Some uncertainties can't be captured in an assumption as we don't have perfect insight. However, effort should be made to identify all possible uncertainties and capture these as assumptions. The assumptions log will convey the boundary of what has been included.

3.4. Sources of analytical uncertainty - Analysis

Undertake appropriate AQA

An additional, but important source of analytical uncertainty is in the analysis itself. With verification and validation of models, good Analytical Quality Assurance (AQA) practices can help identify the restricted uses of analytical outputs and help minimise the possibility of errors. However, mistakes can still be made, so being clear with decision makers about the extent to which analysis has been quality assured can help them understand how far they may rely on analytical results in support of their decision making. Please see the AQuA book for more information.