DRAFT – AVAILABLE FOR COMMENT

Where there is no single Theory of Change: The uses of Decision Tree models

Rick Davies, Original Version: Thursday, 27 December 2012 Revised Wednesday 2nd April 2013

"*At the heart of all major discoveries in the physical sciences is the discovery of novel methods of representation*" Stephen Toulmin (Wikipedia. 2012)

Contents

Abstract: This paper begins by identifying situations where a theory-of-change led approach to evaluation can be difficult, if not impossible. It then introduces the idea of systematic rather than ad hoc data mining and the types of data mining approaches that exist. The rest of the paper then focuses on one data mining method known as Decision Trees, also known as Classification Trees. The merits of Decision Tree models are spelled out and then the processes of constructing Decision Trees are explained. These include the use of computerised algorithms as well as ethnographic methods, using expert inquiry and more participatory processes. The relationships of Decision Tree analyses to related methods are then explored, specifically Qualitative Comparative Analysis (QCA) and Social Network Analysis (SNA). The final section of the paper identifies potential applications of Decision Tree analyses, covering the elicitation of tacit and multiple Theories of Change, the analysis of project generated data and the meta-analysis of data from multiple evaluations. Readers are encouraged to explore these usages.

Included in the list of merits of Decision Tree models is the possibility of differentiating what are necessary and/or sufficient causal conditions and measuring the extent to which a cause is a contributory cause (after John Mayne)

Theories of Change and their limits

Theories of Change (ToC) are in the limelight. This year three reviews have been commissioned in the UK on the uses of Theories of Change, by DFID Evaluation Department (Vogel 2012), Comic Relief (James 2011) and by CARE International (2012). Others have been produced elsewhere (Stein and Valters 2012, Eguren, 2011). There are also websites dedicated to the subject of Theories of Change[1].

An explicit Theory of Change is a great aid to evaluation. At best, it clarifies expectations of outcomes and how they will be achieved, in a way that is evaluable. But ToC have their limits, like all tools. Firstly, many of the Theory of Change representations I have seen have limited capacities for adequately representing complex projects. Funnel and Rogers'(2011) comprehensive discussion of the use of Theory of Change and Logic Models actively warns against introducing too much complexity, including the excess use of feedback loops, because they can make models very difficult to understand. Yet feedback loops are a defining feature of complex systems. Because of this feature complex systems are dynamic, their states change over time. But dynamic models seem to be as rare as hen's teeth, at least in the world of development project evaluation.

The problem lies not only in our limited capacity to represent and understand complex models. Large projects have more stakeholders, generating more perspectives on the expected outcomes of a project, and the way of achieving those outcomes. While participatory planning and evaluation methods can be helpful in identifying areas of consensus about means and ends there are limits to what can be achieved by this approach, especially when there are very different interests at stake. Diversity of views is likely to be a particular problem in projects where there is a significant degree of decentralisation in implementation e.g. in participatory development projects and in portfolios of projects run by different grantees. Advocacy projects would also seem problematic, because they often involve stakeholders with very different views.

There is a third problem that is present, also arising from complexity. Even in the simplest projects with standardised interventions there are many aspects of the context which can affect the outcomes. Pawson and Tilley's (1997) example of the variable results of installing closed circuit cameras for surveillance purposes, in different locations, is a classic example. Where interventions are also varied in character, the number of potential influences on outcomes is greater still. The point to note here is that these influences may not simply act as sole causes, they might also or only be effective in combination with others, a point that will be returned to later in this paper. The number of possible *combinations* of these influential attributes increases exponentially, not arithmetically, as the number of attributes being considered is increased. With 10 attributes there are 2^{10} or 1024 possible *combinations* of these that *might* be associated with significant performance differences. With 20 possibly relevant attributes there are 2^{20} or 1,048,576. The combinatorial space grows very large very quickly. A project's official Theory of Change will represent just *one* of these combinations. In such circumstances it would seem unwise to ignore the rest, even if many would seem rejectable on first sight. Elimination of rival hypotheses is supposed to be part of an evaluator's tool kit for establishing causality claims, but the question is how to do so systematically and comprehensively.

[1] https://www.theoryofchange.org/

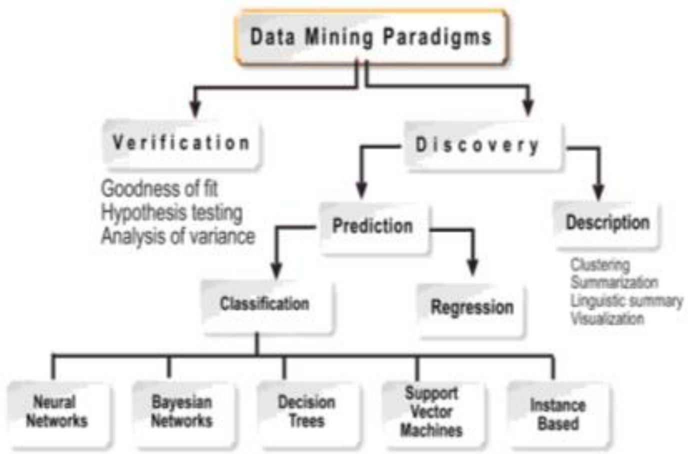
Pritchett et al (2012) argue that many development projects are located in a high dimensional and rugged design space. There are many attributes of the design of a project that need to be set up correctly, if they are to replicate the results of previous projects whose success has been validated with RCTs. Looking at three examples of Conditional Cash Transfer Programs (CCTs) they identify at least 11 specific features that are needed. Small variations in these conditions, many of which are binary rather than continuous, can lead to dramatically different outcomes.

Data mining: ad hoc and systematic

In evaluation and research circles data mining can be seen as a "bad thing", in as much as it appears as an ad hoc search for correlations when perhaps the expected correlations were not found (Backhouse and Morgan 2000, White 2011). It is rightly claimed that such correlations might simply be chance events, with no underlying causal mechanisms at work[2]. But it is equally true that there might be some underlying casual mechanism connecting the correlated events. The fact that the correlations were found by an ad hoc or even random search would not undermine the significance of that finding. The real problem lies in the incomplete and unsystematic nature of ad hoc data mining. There may be other causally linked correlations out there, and they may be more important, but not yet discovered. What is needed is a systematic and comprehensive search process.

In the worlds of business and physical science data mining is seen in more neutral terms. Wikipedia defines data mining as "…the process that attempts to discover patterns in large data sets"[3] As the Wikipedia entry makes clear, the range of applications is enormous and the variety of methods of data mining is considerable. Data mining tools are used by business to analyse consumer behaviour, by finance companies to analyse loan risk, by investors to analyse investment opportunities, by medical researchers for diagnostic purposes, and by many others. There are now a number of major texts on the subject, covering a wide range of approaches[4]. Rokach and Maimon (2008) have produced the following taxonomy:

Figure 1: Taxonomy of data mining methods



Terminology used to describe the discovery & prediction forms of data mining varies according to location of use. In academic environments "machine learning" is the preferred terminology, whereas in commercial

[2] Leaving aside the additional risk that there may be selective reporting results found by ad hoc search
[3] http://en.wikipedia.org/wiki/Data_mining
[4] See Amazon books search for "Data mining"

environments it is called “predictive analytics”. The construction and use of Decision Tree and other models is commonly called “predictive modelling”. Predictive modelling has been defined as “the process by which a model is created or chosen to try to best predict the probability of an outcome”[5]

The focus in this paper is on Decision Trees only, because of their recognised advantages. These have been summarised as follows, and will be explored in more detail later in this paper:

- People are able to understand decision tree models after a brief explanation.
- No assumptions are made about the relationships between the data (e.g. normal distributions, linear relationships)
- Data preparation for a decision tree is basic or unnecessary.
- It is possible to validate a model using simple statistical tests.

The next part of this paper looks at Decision Trees models, as a particular kind of summary representations of knowledge. This will then be followed by an examination of the methods used to generate these models, including Decision Tree algorithms that can be embodied in software.

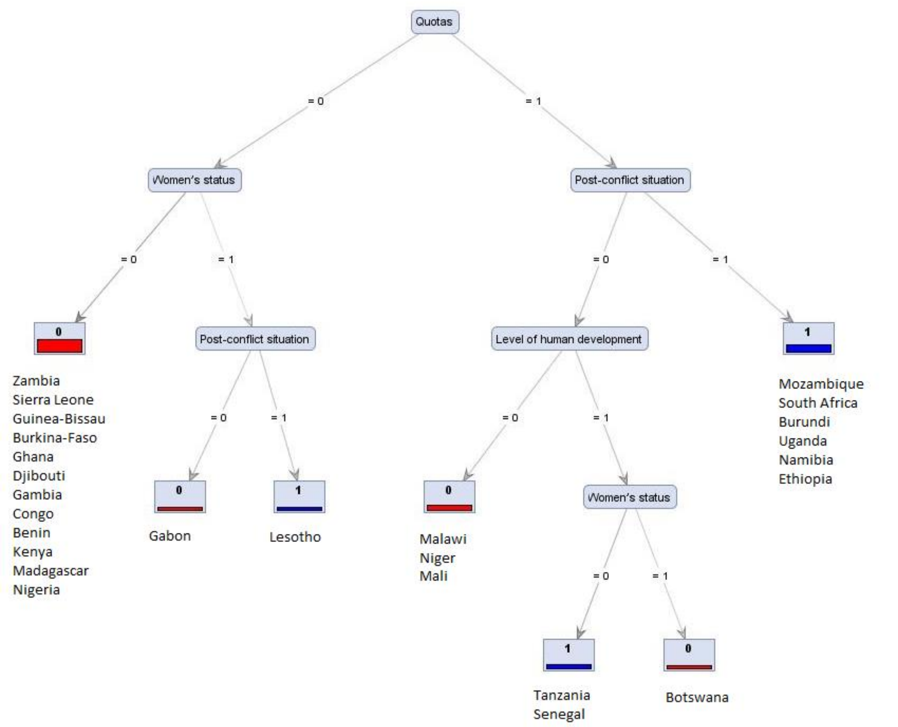
Decision Trees as models

A Decision Tree is a kind of model, a useful simplification of reality. It can be used to summarise how different combinations of conditions are associated with different kinds of outcomes. Applied to an existing set of data about what conditions are associated with what outcomes, it provides a summary *classification*. When the same classification is applied to a new but comparable set of data it provides *predictions* about what outcomes are associated with what combinations of conditions.

Figure 2 below is an example of a Decision Tree, which classifies 26 different African countries according to whether they have a high proportion of women in Parliament or not. The contents of this Decision Tree is based on data and analysis available in a paper on "Women’s Representation in Parliament: A Qualitative Comparative Analysis" by Krook (2010). Please suspend your judgement on the validity of this analysis for the time being, and focus on how the results have been represented. Validity issues associated with this kind of analysis will be addressed later in this paper.

[5] See Wikipedia: http://en.wikipedia.org/wiki/Predictive_modelling

Figure 2: Example Decision Tree



Key: 0 = attribute not present, 1 = attribute present. Red = countries with low levels of women in parliament, Blue = countries with high levels.

A Decision Tree looks like an inverted tree. It has a root node, internal nodes and terminal nodes. The root and internal nodes contain *test conditions* used to separate *cases* that have different attributes. In the above example, the cases are listed under each of the terminal (leaf) nodes.

So, the six countries on the right all fulfil two conditions. The first test condition, at the top of the tree, asks for a given country under examination whether there are quotas for the proportion of members of parliament that must be women. If the answer is yes (=1) than the next test condition asks if those countries are in a post conflict situation. Krook's data set shows that six of those countries are in post conflict situations. All of these have a high proportion of women in parliament.

On the left twelve countries fail two conditions. They do not have quotas for women in parliament and women's status in those societies is low. All twelve countries have a low proportion of women in parliament.

In the middle there are five other configurations of conditions, variously associated with high or low levels of women in parliament. Here a configuration is a specific branch of the tree, with a set of cases as the leaves. Note that configurations can include both the presence and absence of different conditions. For example, Malawi, Niger and Malawi all have quotas for women, but they are not in post conflict situations and they do not measure highly on the UNDP human development index.

Decision Trees are not limited to binary branching structures. It is possible for test conditions to differentiate three or more different types of cases. These are known as multivariate splits.

Decision Trees have a number of merits as representations:

Decision Trees are able to describe multiple means of achieving the same kind of outcome.

This is a property known as equifinality[6]. As shown in Figure 1 there are three different configurations of conditions that are associated with high levels of women in parliament. The same multiplicity of configurations is common in real life. For example, there might be a portfolio of projects funded by a grantee, all aiming to achieve the same outcome e.g. reduced maternal mortality, but different projects may involve different combinations of interventions. On the other hand, at the level of an individual projects although there may be one expected outcome, the social and physical conditions present in different locations within the project area may differ, so the interventions may need to be locally customised. In participatory development projects different communities may seek to reach the same objective of poverty alleviation by different means.

Decision Trees are representations that can acknowledge causal diversity. In contrast, most Results Chain type representations such as LogFrames present one package of activities as a sufficient means of achieving the desired outcome, even though in practice different combinations of activities may be carried out in different locations or with different groups. Network models can provide more options in as much as they describe multiple causal pathways.

This capacity to represent causal diversity is consistent with the emphasis in some schools of evaluation and analysis on identifying the different *configurations* of conditions that can lead to a desired outcome. In Realist Evaluation these are in the form of different combinations of Context and Mechanism, leading to different Outcomes. In Qualitative Comparative Analysis (QCA) multiple explanatory rules typically need to be identified to account for all observed outcomes.

Decision Trees are also able to discriminate between symmetric and asymmetric causal relationships (Goertz and Mahoney 2012). In Figure 2 the causes of low levels of women's representation in parliament found in some countries are not simply the absence of the causes of high levels of women's representation found in other countries. In analyses of other data sets the causal factors may or may not be found to be symmetric.

Decision Trees can use widely available data

Decision Trees are about classification of cases based on their attributes and whether certain attributes are associated with a prescribed condition or not. As such they can make use of categorical (i.e. nominal) data, which is widely available. Such data can be generated by participatory evaluation processes, expert judgements or the partitioning of more sophisticated quantitative measures using ordinal, ratio or interval scale data. There is no requirement that the distribution of categories follow any kind of regular distribution i.e. a normal curve or otherwise. Nor do assumptions need to be made about the kind of relationships between the categories (e.g. independence or linear relationships). In addition, Decision Trees can also be produced using ordinal, interval or ratio scale data[7] and using fuzzy sets[8].

Decision Trees are evaluable

When used for classification proposes Decision Trees vary in their *discriminatory power.* Where a given branch leads to outcomes of one kind only (as in Figure 2), rather than say 90% or 70% of one kind, these are said to have higher *discriminatory power*.

---

[6] http://en.wikipedia.org/wiki/Equifinality

[7] For example, BigML online Decision Tree software

[8] See Google Scholar search results for 2012 "fuzzy set decision trees"

Decision Trees with few branches and few conditions within these branches have *simplicity*, which aids interpretation of the Decision Tree.

When used for prediction purposes Decision Trees vary in their *accuracy*. For example, other countries in Africa could be subject to the same set of test conditions present in the Figure 2 Decision Tree. It is possible that there may be two other countries that have quotas and which are in post conflict situations, yet the percentage of women in their parliaments is low. In this situation we could say that in the configuration to the right there is 75% accuracy (6 of 8 cases in the configuration are correctly classified).

Further distinctions can also be made between *sensitivity* (i.e. True Positives/Actual Positives) and *specificity* (i.e. True Negatives/Actual Negatives).

Decision Trees can also vary in *stability*, i.e. their predictive accuracy over time.

Moore et al (2001) suggest that the most desirable Decision Trees are:

1. Accurate (low generalization error rates)
2. Parsimonious (representing and generalizing the relationships succinctly)
3. Non-trivial (producing interesting results)
4. Feasible (time and resources)
5. Transparent and interpretable (providing high level representations of and insights into the data relationships, regularities, or trends)

Decision Trees pay attention to internal and external validity

Decision Trees are typically developed on the basis of an examination of a set of data about cases where both their conditions and their outcomes are known (known as training cases). They can be assessed in terms of the accuracy with which they categorise the known cases. The same Decision Tree can also be used to predict expected outcomes when applied to a new set of cases with comparable kinds of attributes (known as test cases). For example, Ryan and Bernard (2006) developed a Decision Tree that was 90% accurate in its ability to correctly classify recycling behaviour of 70 informants in the USA. When it was applied to a nationwide sample of respondents it still managed to achieve an 84% level of accuracy.

In the Decision Tree literature it is recognised that there can be a trade-off between ability to accurately classify the training cases and accurately predict the outcomes in test cases. Decision Trees that are highly accurate descriptions of training cases may fail to accurately classify test cases. This risk is known as "over-fitting". The solution is to "prune" the Decision Tree i.e. remove some of the lower level conditions and simplify the model[9], at the cost of its accuracy in describing the training cases.

Decision Trees provide a modest form of counterfactual

Goertz and Mahoney (2012) differentiate between within-case and between-cases approaches to counterfactual analysis. A within-case approach involves the development of potentially testable conjectures about what would have happened if a condition X was not present. A between-case approach involves comparisons with other cases. In its extreme form the other cases are controls which are identical except for the presence of a condition X. Alternately; they argue that "a plausible counterfactual in qualitative research is one where there are cases in the dataset that are similar to the counterfactual being proposed". This approach seems a more realistic approach when cases vary on multiple attributes, not just the presence/absence of a condition X.

[9] Especially those conditions that differentiate a small number of cases.

Comparable cases can be identified within a Decision Tree model, and in the underlying data sets used to construct them. The degree of similarity between two cases can be described by their number of shared attributes. For example, in Figure 1 Tanzania, Senegal and Botswana have three conditions in common, but Tanzania and Senegal differ from Botswana in that women's status is lower in their countries. This single difference is associated with a difference in outcomes. If we then look at the underlying data set (Figure 6 below) we can see there are no other differences between them. There may of course be other important differences outside the current data set, which could be investigated.

Decision Trees can enable the differentiation of different types of causes.

John Mayne (2012) is well-known for championing the need to differentiate causal contribution from causal attribution[10]. However, in a recent and associated paper Michael Patton (2012) reported concerns that a contribution analysis will always find a contribution of some kind and that the concept of contribution is so broad that that "any finding of *no contribution* is highly unlikely"[11]. To avoid this problem it would be useful to be able to differentiate the degree to which a condition is a contributing cause. Being able to do so should be very useful for evaluation purposes. Decision Trees provide a means of doing so, which will be explained below.

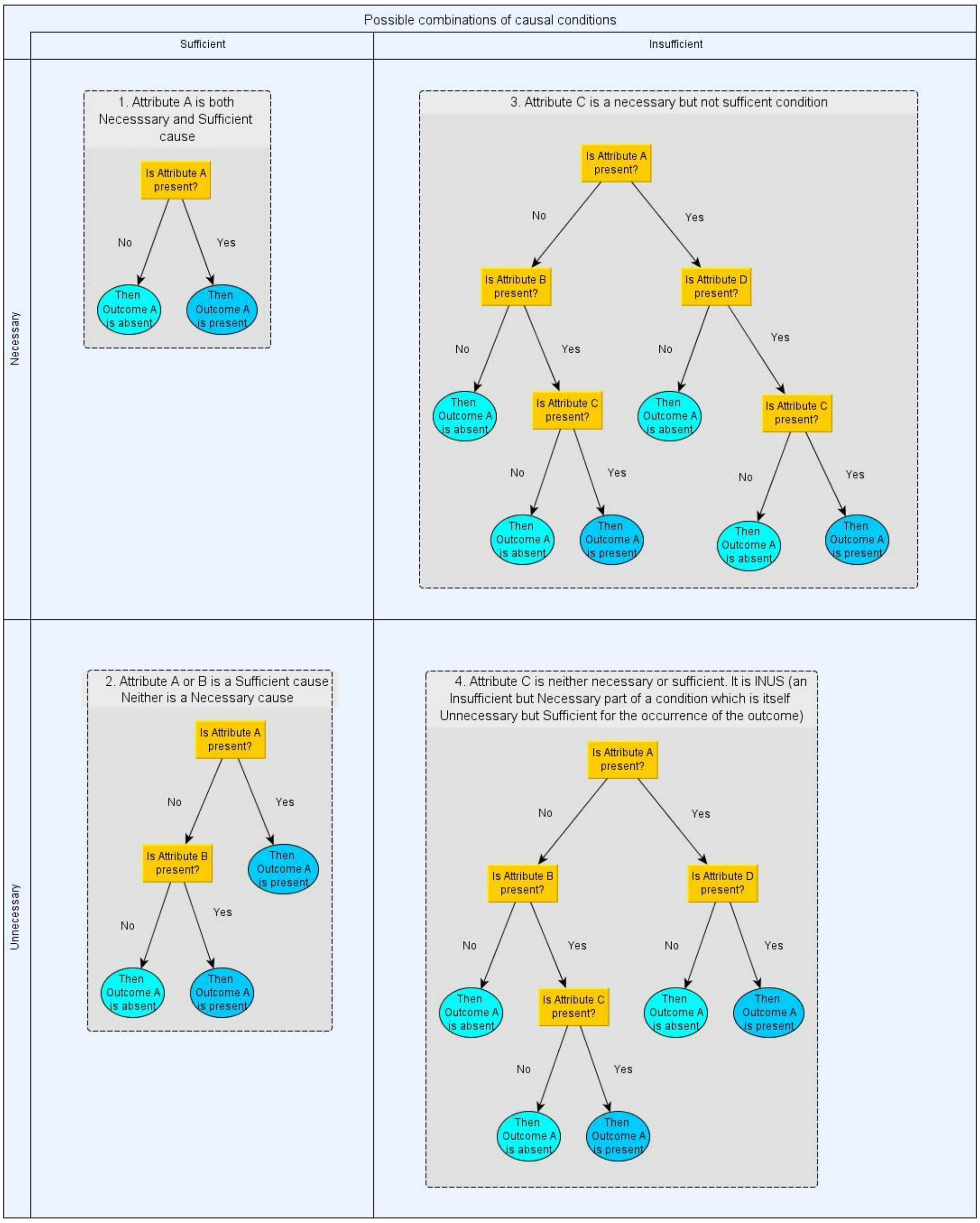
As pointed out by Mayne (2012) and others (Stern et al. 2012), the literature on causality differentiates between conditions which may be a necessary cause or sufficient cause, or a combination of these. The difference between these kinds of causes can be visualised in the structures of a Decisions Tree, as shown in Figure 2 below.

[For the temporary purposes of this exposition, assume that the associations shown in the Decision Tree are in fact causal associations. This assumption will be revisited below]

[10] Some would argue that this is largely a rhetorical difference, but even if so it does usefully emphasise the idea of multiple causes and influences.

[11] John Mayne has subsequently commented "It is probably true that in a prospective sense, one could often put together a 'theory' linking the intervention and a desired effect. But in terms of credibly demonstrating a contributory cause, this is a much more challenging undertaking, as other articles in the Issue show. Indeed, my concern is rather that is may be quite difficult in many cases to demonstrate a contribution. As you have argued elsewhere, theories of change can be rather extended and complicated, and showing that all the links have worked and that there are not other reasonable explanations is very demanding. That, I think is the challenge of contribution analysis, not that a contribution can be readily shown in most cases. It can't, other than in a hand waving sense" My recent experience of reading reports of DFID projects in India is that claims of contribution are casually made all too often, and this is the more common of the two possible problems.

Figure 3: A visualisation of the possible combinations of necessary and sufficient conditions



If an organisation is seeking to claim maximum impact it could be argued that they would order these types of conditions in a hierarchy of importance, as follows[12]:

1. Necessary and sufficient causes – without the intervention nothing would have happened
2. Sufficient causes – the intervention was sufficient by itself
3. Necessary but not sufficient causes – the intervention was needed but other conditions were also needed.

[12] However if it was seeking to claim maximum sustainability the ordering might be in reverse

4. Neither necessary nor sufficient causes – other kinds of interventions could have produced the same outcome.

In the last category there are two sub-categories. One, shown in Figure 3 above, are conditions that are insufficient but necessary part of a configuration that is not necessary but sufficient to cause an outcome (known as INUS conditions[13]). In Figure 1 the existence of quotas is an INUS condition. The other sub-category is the conditions that do not even qualify as a necessary part of such a configuration. As such they would not even appear in the structure of a Decision Tree, because they do not enable distinctions between outcomes (e.g. high and low levels of women in parliament)[14]. This potential to differentiate degrees of contribution should allay Patton's concerns.

Philosophers have argued that in many situations being examined we are looking at the fourth category, INUS conditions[15]. This would seem to be the case with most development interventions. There is usually more than one way of addressing a problem and more than one agency that could do so. Exceptions might be found in humanitarian emergency work. For example, where a helicopter delivery of emergency assistance is needed for communities in isolated mountain areas following an earthquake. That might qualify as necessary and sufficient, at least for some purposes.

Within the more common INUS situations further distinctions can be made about the relative importance of a given condition. Looking back at Figure 2 we can see that the presence of quotas is a contributory cause in seven of the eight cases where there was a high level of women's representation in parliament, whereas high level of women's status was a contributory cause in only one of the eight cases. More generally, it seems that the higher up the tree (i.e. nearer to the root node) the more important is the role of the condition, because it will be part of a greater number of configurations[16].

Caveats

Decision Trees are about associations between conditions and outcomes, and associations are not by themselves evidence of causation. There also needs to be some evidence of, or plausible argument for, the existence of a causal mechanism that leads the associated conditions to generate the observed outcome. Without this, there is a risk that the association is spurious, a coincidental event arising perhaps from some other shared influence. A good claim of causal attribution requires the combination of some form of co-variation plus mechanism. One without the other is not sufficient. This necessity is recognised in the approach taken by 3ie's with the funding of RCTs, which encourages the use of a Theory of Change to accompany and support the statistical evidence generated by RCTs (White 2012).

Theories about change can inform two stages in the development and use of Decision Trees. At the beginning they can inform the choice of possibly relevant test conditions that may form a useful Decision Tree. Different stakeholders may have different views of what attributes of an intervention will make a difference to the expected outcome. As will be shown below, a range of such views can be accommodated and their relevance tested, during the development of a Decision Tree.

Once different configurations of attributes have been identified as being associated with specific outcomes then theories can also help identify the mechanisms connecting the attributes in the configuration. Because there may be multiple configurations multiple theories may be useful.

---

[13] INUS = Insufficient but Necessary part of a configuration that is Unnecessary but Sufficient

[14] Decision Trees typically make use of only a subset of all the attributes in a dataset

[15] See http://en.wikipedia.org/wiki/Causality

[16] However this will not always be the case The exception being a condition that might appear in the lower end of multiple branches.

There are no absolute requirements for an adequate causal mechanism. More detailed /fine grained descriptions are better than ones less so because they are more open to disproof. A mechanism that includes specific links between the component parts is preferred to one without for the same reason[17].

The recognition of these roles for theories of change does not contradict the position taken at the beginning of the paper, which was about the limits of the use of a single Theory of Change approach.

Counter-caveats

Valid *explanations* of causal processes behind the associations found in a configuration may not always be needed. Decision Trees and other methods (e.g. artificial neural networks), may generate accurate *predictions* which are useful in themselves, without any knowledge of the underlying causal processes. These are known as "black box" models. Accurate predictions of public behaviour in response to immunisation campaigns and to the provision of other government services could make a substantial difference to the design of such services and thus their subsequent uptake. Not surprisingly, there is significant on-going research on the use of Decision Trees to predict stock market behaviour[18]. Those involved are not seeking to understand and subsequently influence stock market behaviour, just to profit from its behaviour as it emerges. On the other hand valid explanations are useful when activities are being designed with the intention of *producing* the desired outcome. For example, changing people's health seeking behaviour.

The construction of Decision Trees

There are at least two means of constructing Decision Trees:

- By Decision Tree data mining algorithms
- By ethnographic and participatory inquiry

There are two other methods which are similar in purpose but which won't be discussed here:

- Software used for the production of cladograms[19], which are tree structures showing the relationships between different species. Classifications of species reflect the most parsimonious combination of their attributes. Here there is no "training" set available with cases where the relationship between attributes and outcomes is known. The process here is more akin to clustering, as given in Figure 1 above.
- Decision Trees as used for management purposes, which have probabilities assigned to each branch rather than test conditions[20]. Outcomes are given financial values and the values of different branches reflect the sum of financial values x probabilities. Here the results of interest are the values of different branches of trees, each of which represent different scenarios or strategies.

Decision Tree algorithms

An algorithm is a procedure spelling out a series of steps that will generate an expected outcome. Algorithms embodied in software can be applied to large numbers of cases in a small period of time. Decision Tree software usually contain a number of alternate algorithms for generating Decision Trees. These algorithms contain instructions for appropriate "splitting" of branches and appropriate "pruning" of the completed tree, and associated methods for assessing Decision Tree performance.

---

[17] It is interesting to note that at the level of explanatory mechanism we seem to need the opposite of Occam 's razor, because short chains of events would be harder to disprove than longer ones.

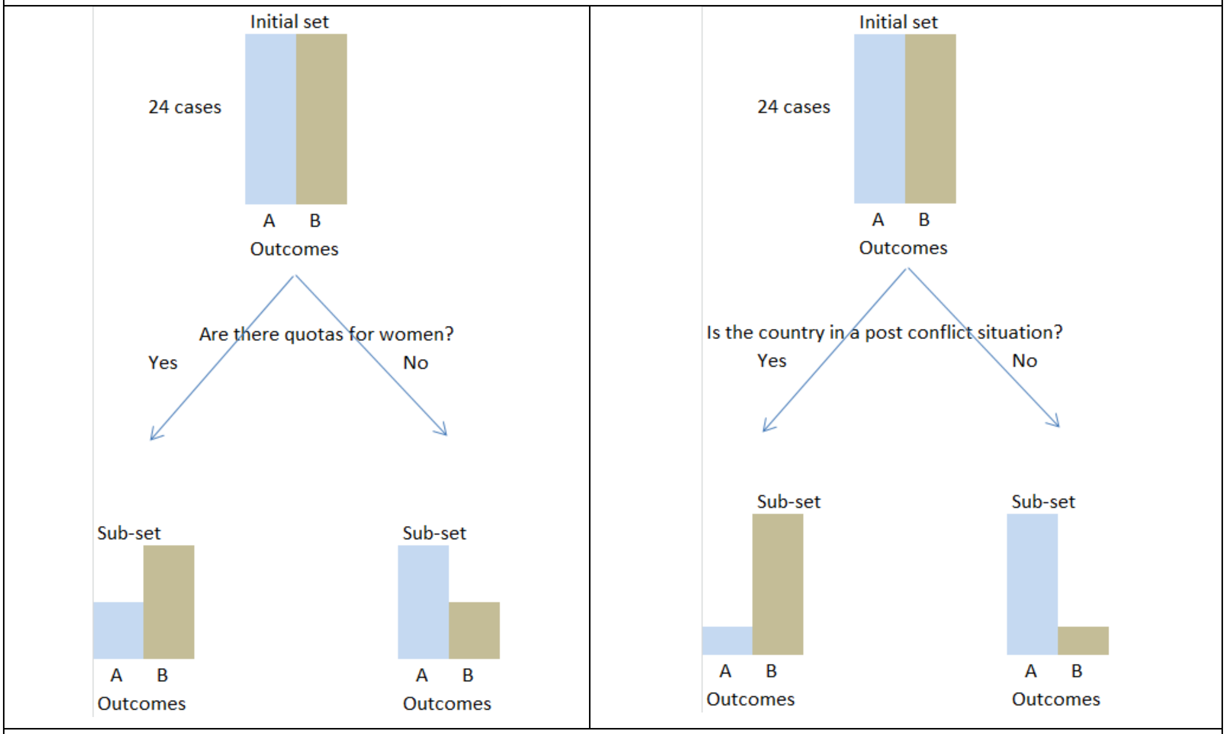
[18] See Google Scholar search results on "decision tree" + "stock market"

[19] See http://en.wikipedia.org/wiki/Cladistics

[20] See http://en.wikipedia.org/wiki/Decision_tree

The core idea behind the construction of a Decision Tree is the progressive reduction of diversity in a collection of cases, from the diverse membership of the initial training set, down to a number of individual sets (the "leaves"), each of which contain a specific kind of case. This is done by a systematic search for a test condition (e.g. "Does the country have quotas for women in parliament?") which most effectively splits all cases into two groups, in which each contain a more homogenous set of cases (described as "purity"). In Figure 3 below there are two imagined attempts to split all the cases in a training set, using two different test conditions. The second test condition is more effective because it has led to a bigger increase in homogeneity with each of the sub-sets of cases, when compared to the initial training set[21].

Figure 4: Comparison of the effectiveness of two test conditions in increasing purity of cases.



Data is fictional

Decision Trees algorithms can use a variety of tests to automatically identify which test conditions provide the most effective split. The simplest to understand is a Chi-Square test[22]. This can be carried out for the results of both tests above. The tests are shown below.

Figure 5: Chi Square tests for results of the two tests

		Quotas?	
		Yes	No
Outcomes	A	2	4
	B	4	2

Chi squared equals 1.333 with 3 degrees of freedom.

		Post conflict?	
		Yes	No
Outcomes	A	1	5
	B	5	1

Chi squared equals 5.333 with 3 degrees of freedom.

[21] See Scott Page's video lesson on Categorical Models at https://class.coursera.org/modelthinking/lecture/38 In this video Page explains in simple terms how categorising a set of items into two sub-sets of items can reduce variation amongst items, a process which is the basis of the design of Decision Tree algorithms.

[22] Others include the Gini Coefficient and entropy measures

Once the best split is identified the same procedure is re-iterated: used again to split each of the two sub-groups into even more homogenous sub-sub-groups. This process is repeated until all sub-groups are completely homogenous or the size of the sub-group of cases has reached the lowest allowable limit (to prevent over-fitting).

Data sets

Data for analysis by Decision Tree software is typically presented in a simple matrix form, with cases presented row by row, and their attributes presented column by column, with the outcome of interest in the last column. The example data set in Figure 6 has undergone some pre-processing, with conversion of the numerical measures in the fourth and sixth columns to binary measures[23].

Figure 6: The Krook dataset used to generate the Decision Tree in Figure 2.

Country	Electoral system	Quotas	Women's status	Level of human development	Post-conflict situation	% Women in national parliament
Mozambique	1	1	0	0	1	1
South Africa	1	1	1	1	1	1
Burundi	1	1	0	0	1	1
Tanzania	0	1	0	1	0	1
Uganda	0	1	1	1	1	1
Namibia	1	1	1	1	1	1
Lesotho	0	0	1	1	1	1
Senegal	0	1	0	1	0	1
Ethiopia	0	1	0	0	1	1
Zambia	0	0	0	1	0	0
Sierra Leone	1	0	0	0	1	0
Guinea-Bissau	1	0	0	0	1	0
Malawi	0	1	1	0	0	0
Gabon	0	0	1	1	0	0
Niger	0	1	0	0	0	0
Burkina Faso	1	0	0	0	1	0
Botswana	0	1	1	1	0	0
Ghana	0	0	0	1	0	0
Djibouti	0	0	0	1	1	0
Mali	0	1	0	0	0	0
Gambia	0	0	0	1	0	0
Congo	0	0	0	1	1	0
Benin	1	0	0	1	0	0
Kenya	0	0	0	1	0	0
Madagascar	0	0	0	1	0	0
Nigeria	0	0	0	1	0	0

Risks and limitations

There are at least four risks:

- The number of cases may be so small that external validity will be poor. Internal validity (as in accuracy of the classification of the training cases) may still be high, and of value in itself. External validity will be enhanced by the inclusion of a diversity of cases in the training set[24].
- The attributes may be poorly chosen, in the sense that there was no likelihood of any meaningful association between them, so the results that are found are obviously spurious.

---

[23] Many Decision Tree software packages do not require numerical measures to be converted to binary form.

[24] The figure 6 data set contains 12 of 64 [i.e. 2^6] possible combinations of six attributes. The addition of the 31 other countries in Africa could increase the diversity of combinations.

- The reliability of the assessments made of the attributes and outcome may be low, introducing "noise" and generating inaccurate classifications and poor predictions.
- There may be too many missing observations. While Decision Tree algorithms can cope with some missing data there are limits.

There is a large literature on the performance and merits of different Decision Tree algorithms, which goes into much more detail. Somewhat surprisingly, and fortunately, it seems that results are relatively insensitive to differences in splitting and pruning procedures. "Ensemble" methods, such as Random Forests[25], that generate and analyse multiple Decision Trees generated from one data set have been found to be more accurate, but the results present "readability problems for most people. Decision Tree algorithms that can work with fuzzy set data have been shown to perform better by a modest margin on a number of standard test data sets (Sachdeva, Hanmandlu, and Kumar 2012)[26]. Careful selection of cases for inclusion within a training set ("instance selection), which is designed to maximise diversity, has also been shown to be helpful.

The software available

I have tried out the following packages:

- dTree – A free package that runs on java (easily installed on most PCs). Easy to use, and recommended. Can only use nominal data.
- BigML - An online service. Unconventional in structure but easy to use and modestly priced. Can use interval and ratio scale data
- RapidMiner – A sophisticated open source data mining suite, but demanding to learn the basics.
- XL Miner – An Excel plug in, easy to use but expensive. Free trial.
- GA Tree – A free version available, that uses genetic algorithms to find best fitting trees
- Rattle: A data mining "windows" interface for R (an open source stats programming language). Free and comprehensive. Steep learning curve if no prior knowledge of R.

Lists of free and commercial software packages are available online:

- http://www.kdnuggets.com/software/classification-decision-tree.html
- http://www.the-data-mine.com/Software/MostPopularDataMiningSoftware

See also: "A Survey of Open Source Data Mining Systems" by Chen (Chen, Williams, and Xu).

Manual construction of Decision Trees

Ethnographic inquiry

The classic description of the ethnographic approach is Gladwin's (1989) "Ethnographic Decision Tree Modelling". Prior to that publication Gladwin had used Decision Trees to develop models of farmers agricultural practices in Africa and the Americas. The strength of her approach is in the ethnographic attitude, oriented towards identifying participants own decision making criteria, rather than a researchers more etic view. Gladwin's early work has been followed by applications in many other areas[27], including the following:

---

[25] http://en.wikipedia.org/wiki/Random_forest

[26] Fuzzy category membership values can range anywhere from 0.00 to 1.00, rather than being either 0 or 1

[27] See a pre-2000 list of references here: http://www.analytictech.com/mb870/Handouts/edm_references.htm

- Health: Mother's breastfeeding behaviour, patients' choice of heart disease treatments, skin cancer patients use of sun protection methods, mothers' choice of childbirth locations, drug users choices re needle sharing, carers response to sick children, treatment seeking by stroke patients
- Technology: Adults choices of information technology, security decision making in airports, students use of weblogs,
- Agriculture: Farmers adoption of organic farming practices, farmers choice of land management practices, farmers use of credit and fertiliser, farmers choices about tree planting, farmers adoption of new sheep breeds
- Business: Managers decision making, consumers choices about recycling

Here the "cases" are individuals, rather than groups, organisations or states. The outcome of interest is behaviour, the choices people made about medical services or treatments, the uses of technology, and various farming practices. The attributes of these include their own resources, their preferences, and their knowledge of the options available and aspects of the social and economic context.

Gladwin outlines the following steps in developing an ethnographic decision tree model (EDTM):

1. Identify what decision to be examined, the kinds of outcomes of interest
2. Identify the range of alternatives to the decision. These might be binary (yes/no) or multiple choice (referred to as multivariate splits above)
3. Find an informant and carry out an ethnographic interview (as in Spradley 1979), to learn about the cultural scene from the informant's (emic) perspective.
4. Follow up with participant observation of informant(s) carrying out the activities of interest. E.g. farmers using fertiliser.
5. Identify a sample of people to interview about their decision making, including a balanced number of those who decide to do and not to do the activity of interest. While diversity needs to be maximised, Gladwin suggests an upper limit of 25. Though others (Ryan and Bernard 2006) have used up to 70.
6. Discover decision criteria in use, by:
    - Look for contrasts over decision makers, over space or locations (with one decision maker) or over time (with one decision maker)
    - Elicit the criterion by asking "Why did person 1 do X but person 2 do Y?", or "Why did you do X when you were here, but Y when you were there?", or "Why did you do X then but Y later on?"
    - Make first draft of a Decision Tree based on the first interview, as an aid to the subsequent interviews.
7. Build a composite Decision Tree for the group, from the individual Trees, by either of these methods[28]:
    - Building up a composite model, step by step, after each interview.
    - Building multiple individual models then creating one aggregate model at the end.

The aggregation process needs to combine all the informants' criteria "in a logical fashion while preserving the ethnographic validity of each individual decision model". Logical refers to the sequence of decision making criteria making sense e.g. they might be expected to be applied in that sequence in real life. This is not a performance criteria used by Decision Tree algorithms, though it could aid comprehension of the final

[28] Gladwin describes this process as two steps (7 and 8) for reasons which are not clear

Decision Tree[29]. Ethnographic validity implies minimising use of generalisation in place of original actors' descriptions, and ensuring that each participant's criteria is still used in the final version of the Decision Tree and leads to the same outcome[30].

Gladwin distinguishes Decision Tree models from verbal descriptions people's behaviours by their testability, an advantage of Decision Trees noted earlier in this paper. She outlines seven steps in the testing process:

1. Make up a formal questionnaire, using each decision criterion as a question. The answers by the respondent will be yes or no
2. Identify a sample of respondents to test the Decision Tree model
3. Identify what decisions the respondent actually made, before asking how each criterion applied to them.
4. During the interviews note if and where the model is failing i.e. outcomes don't occur as predicted. At the end of the interview seek out additional criteria by contrasting the conditions that suggested one likely outcome with the unexpected outcome.
5. At the end of all interviews calculate the success rate in the model as a whole (i.e. the proportion of all decisions that are correctly predicted). Gladwin suggests that "If the decision model successfully predicts 85-90% of the choices of individuals in the group it is assumed to be an adequate model for that group of individuals". The basis for this performance criterion is not clear, it could be argued that it depends on the kinds of behaviour being modelled. With models of stock market performance a 55% success rate would still be profitable, whereas with models of disease diagnosis much higher levels of success would be needed[31].
6. Adjust the design of the model, based on participants' feedback about errors in the model, to generate a revised model. Improvements can be made by rephrasing decision criteria, adding new criteria, or relocating the criteria within the Decision Tree.
7. Test the revised model with the test sample data, and compare results with the initial model. If new criteria have been added or old ones substantially changed, then testing will be needed with a new sample of respondents. This is because the model has become a descriptive and not predictive model because criteria have been modified to best fit the first test sample.

An alternative ethnographic approach using card sorting

Card sorting exercises are one of the more common methods of ethnographic inquiry (Harloff and Coxon 2005). One card sorting method, known as Hierarchical Card Sorting, can be used to elicit participants' classifications of entities (people, places or events) in the form of a nested classification i.e. a tree structure (Davies 1996).

Figure 3 shows a classification of 30 African and Asian countries, in the form of a tree structure, as seen by a sub-group of staff in a bilateral aid agency. The contents were generated by a Hierarchical Card Sorting process. The yellow square nodes are the test conditions, the green round nodes are the outcomes, the countries thus classified. Their number is shown here but not their individual names[32].

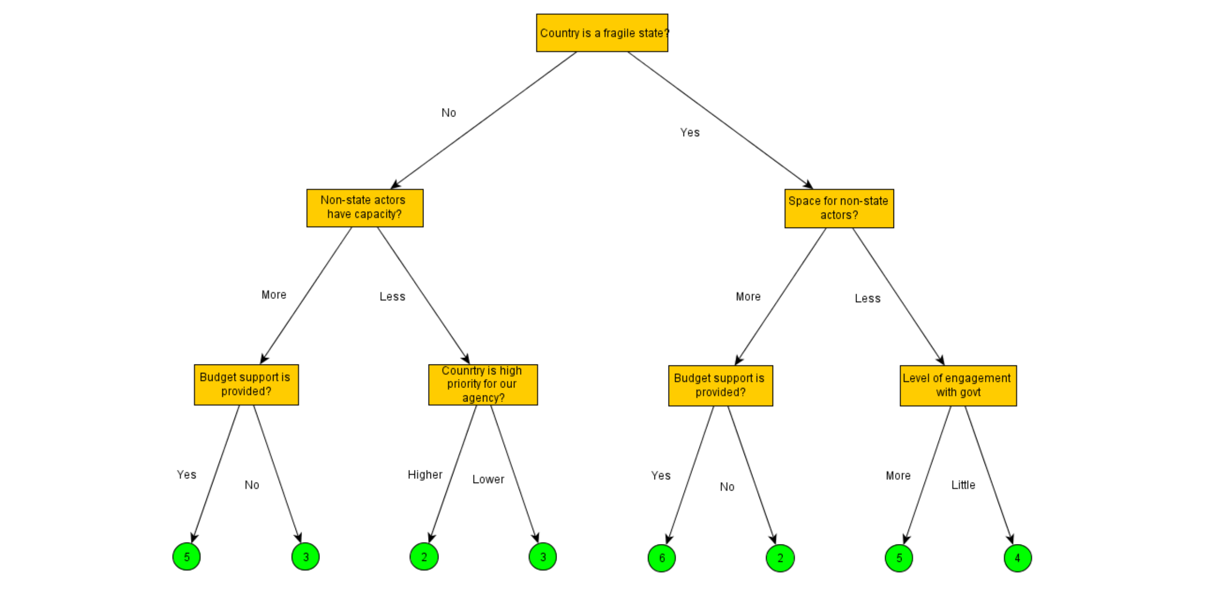
This classification is not yet a Decision Tree of the kind described above, because we don't know the outcomes associated with each branch. But this gap can easily be filled by asking the same participants to

[29] Another approach mentioned by Gladwin is to ensure all outcomes of one side of the tree are of one kind, and all of the other kind are on the other side.
[30] Though possibly not familiar to Gladwin, the splitting methods used by computerised Decision Tree algorithms could also be used to decide which decision criteria to use in a given part of the tree .
[31] Because the cost of failures would be incurred by patients in ways that could not be subsequently redressed.
[32] To anonymise the data source.

choose which groups of countries (in each of the "leaves") they thought were doing better on a performance criteria of interest to them, starting with the distinction at the top of the tree (Country is a fragile state?), and repeating the same question for each sub-group of countries, down the branches below. The results will be a complete rank ordering of all eight groups of countries (but not of the countries within each "leaf"). For the sake of illustration, the branches in Figure 7 have been ordered to reflect a possible result had this extra step been taken. The most "successful country programmes" are to the left and least successful to the right.

This kind of Decision Tree model can be tested in three ways. Firstly, the status of the countries that have been classified as higher or lower in performance can be compared to independent measures of the same performance criteria, if such measures can be identified. This would be a test of *internal or construct validity*. Secondly, other staff in the same organisation could be asked to classify the same set of countries, by applying the Decision Tree that has been constructed by the first group (with outcomes hidden). This would be testing the *reliability* of judgements made across participants. Thirdly, the original participants could be asked to classify a new set of countries, using the criteria embedded in the Decision Tree they have created with the first set of countries. When compared to any independent outcome measures this would be a test of *external validity*, the ability of their model to be generalised to other cases.

Figure 7: Decision tree structure derived from a card sorting exercise



Relationships to some other methods

Qualitative Comparative Analysis

To quote Wikipedia, "Qualitative Comparative Analysis (QCA) is a technique, developed by Charles Ragin (1987) , for solving the problems that are caused by making causal inferences on the basis of only a small number of cases."

There are important areas of difference and similarity between Decision Tree and QCA. The first similarity concerns the type of data used. QCA uses the same kind of data set as shown in Figure 6 above. This may be preceded by some pre-processing of data, to convert what might be a range of numerical values into two or more categorical judgements. Unlike some Decision Tree algorithms QCA can only work with categorical data, not nominal, ordinal or ratio scale data. There are however versions of both QCA and Decision Tree algorithms that are able to work with fuzzy sets, i.e. values that indicate the degree to which an entity belongs to a category or not. There is also an overlap in the performance measures used by QCA and Decision Tree models. Each QCA expression can be measured in terms of their "consistency" (the percentage of cases they accurately classify) and "coverage" (the percentage of all cases that a QCA expression applies to). The same kind of measures can be applied to each branch of a Decision Tree. Both QCA and Decision Trees can also discriminate between symmetric and asymmetric causes.

The second similarity is in how the nature of the sample of cases can affect the strength of the findings. With both QCA and Decision Tree models a more diverse sample of cases in the training set is likely to strengthen the validity of the findings. A greater diversity of cases increases the likelihood that all possible logical combinations of the attributes of interest that might exist will be available for analysis. This stands in some contrast with experimental approaches where the quality of the analysis is strengthened by ensuring that the control group is as similar as possible to the intervention group[33].

A major difference is in the usability of the results. QCA does not generate Decision Tree structures. Instead it generates association rules that have the best fit with all the observed cases. The attributes and outcomes associated with each case are described in Boolean logic[34]. Because cases typically have some similarities in their package of attributes there is the potential to achieve a reduction in the number of different Boolean logic statements that will adequately describe all cases. This reduction process is done through a minimisation procedure which is part manual and part automated. As shown in Figure 8 below, the results, when expressed in Boolean notation, are not in easily communicable form, and not easily assessed using the kind of performance measures mentioned above. However, the same notation can be manually converted into a more readable Decision Tree, as seen in Figure 2.

Figure 8: Results of Krook's QCA analysis of the data in Figure 6
More women in parliament = *QU * PC + WS * PC + QU * ws * DE* *Less women in parliament* = *qu * ws + WS * pc + de * pc* Clue: in Boolean notation the symbol "+" means OR and the symbol "*" means AND. The letters in upper case refer to conditions present and the letters in lower case refer to conditions absent (quotas, women's status, post-conflict situations, development level)

There is also a difference in the underlying methods of analysis that are used. The methods used are quite different, with QCA comparing the merit of Boolean logic descriptions of whole configurations of attributes, whereas Decision Tree algorithms pay no attention to logic and simply seek to minimise diversity (aka entropy) in each set of cases it deals with. It is described as a "greedy" algorithm because it progressively looks at the next most useful attribute, not a whole set of attributes at a time. However, because they can both work with the same set of data and produce results which are comparable, the two methods can provide a form of triangulation.

In the case of Krook's data, the Decision Tree results are consistent with the QCA results. All the configurations described in the Boolean statements in Figure 8 can be found in the Decision Tree. However,

[33] See Fiss (2009) for a tabulation of maximum attribute numbers desirable for each number of cases, from 10 to 45, based on experimental design with random data matrices

[34] See Wikipedia http://en.wikipedia.org/wiki/Boolean_algebra_%28logic%29

by combining high and low outcomes in the same diagram the Decision Tree manages to make use of fewer attributes in total (13 versus 12), providing a slightly more parsimonious description.

Both methods do not always generate the same result. Recently Fischer (2011) did a QCA analysis of the causes of conflict in policy networks in Switzerland. He found five configurations which were sufficient conditions to distinguish cases of conflict and non-conflict. However, a re-analysis of the same data using a Decision Tree algorithm identified four conditions which were sufficient. Of these three were the same as the QCA analysis. Some of the difference in findings might be explainable by the fact that data set contained only seven of the sixteen possible combinations of conditions. A larger number of cases, perhaps representing a wider variety of configurations of conditions, could help resolve which set of rules was the most useful.

Network Analysis

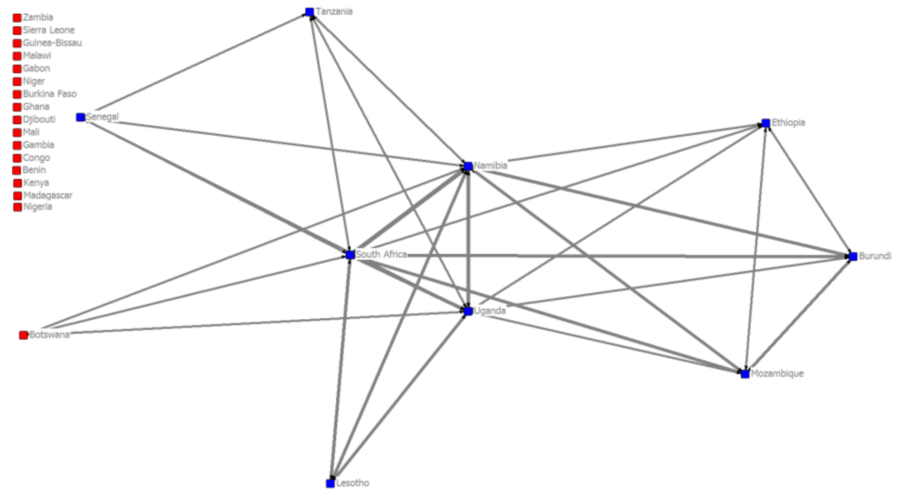
There are many different ways of doing network analysis, as there of doing data mining. Some forms of network analysis can also be seen as a form of data mining. Cluster analysis, as shown in Figure 1 (under Data Mining>Description) is one, because it is a form of pattern identification. Cluster analysis can be carried out with the data in the Figure 6 format, to identify clusters of countries and clusters of attributes. Figure 9 shows a cluster of 10 countries in the Crook data which was found by very simple means when a filter was applied to select all countries that shared three or more attributes in common (but which may vary in content from dyad to dyad)[35]. This cluster contains all the “high levels of women’s participation” countries plus one low level country (Botswana). The existence of Botswana as an outlier suggests that further investigation of its characteristics might be worth investigation – why does it have a low level of participation when it shares many attributes in common with countries with high levels of participation?[36] In other respects the structure of the cluster shares features with the Decision Tree in Figure 2, with Tanzania and Senegal standing out from the others, as does Lesotho.

In this application network analysis does not provide a form of triangulation, because a different kind of output is being produced. But it is providing a different and potentially useful perspective on the same set of data.

[35] Using UCINET&NetDraw, a widely used network analysis package, at https://sites.google.com/site/ucinetsoftware/home

[36] Lesotho may also be worth investigation, standing out as an exception in Figure 2, form all other countries with no quotas a low women’s status

Figure 9: Countries with 3 or more shared attributes

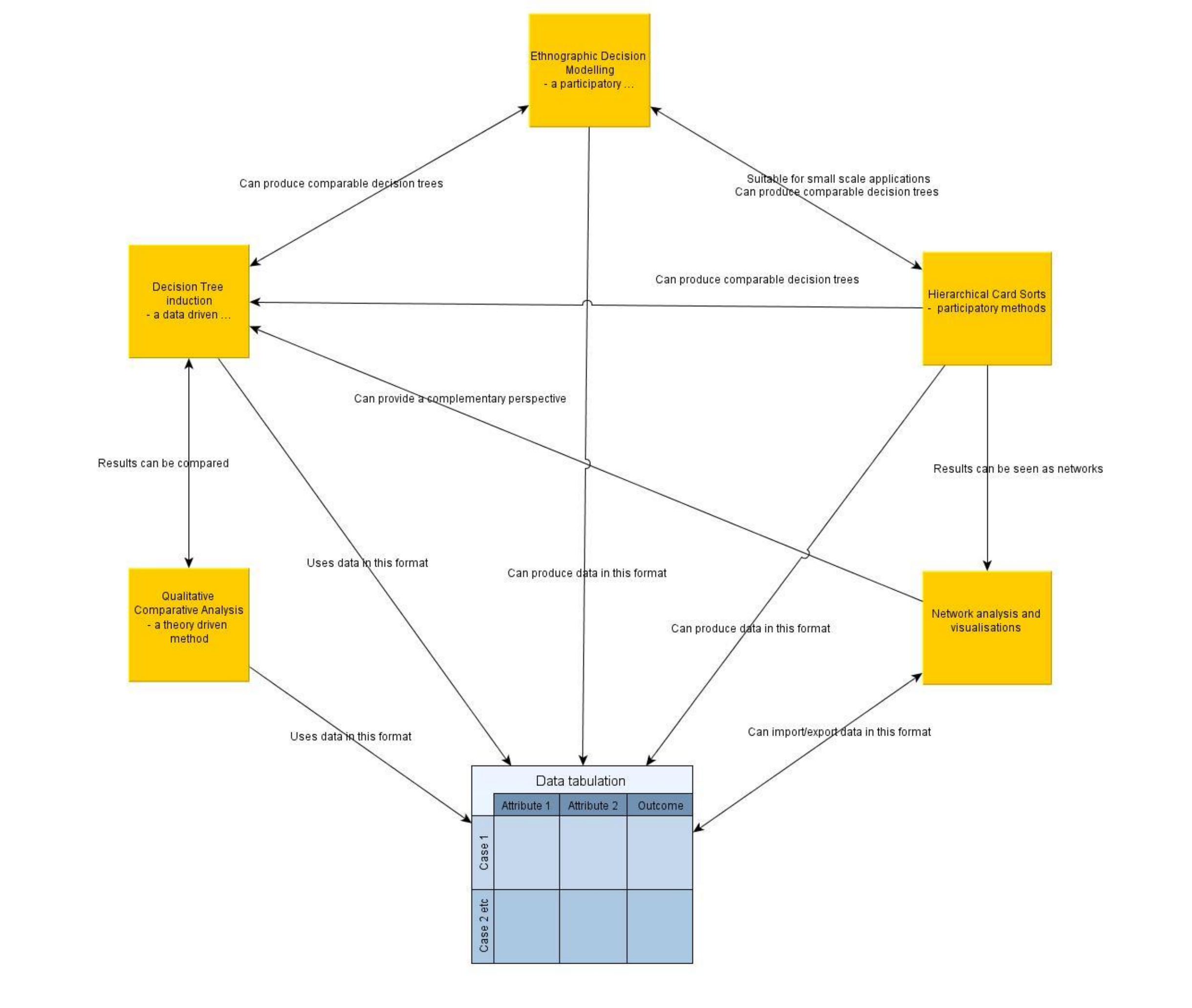


Blue nodes = countries with high levels of women's participation
Red node = countries with low levels
Line thickness = number of shared attributes. More = thicker
Note: Distance between the countries is irrelevant, what matters is differences in the structure of the connections between countries

In summary…

While Decision Trees can be used as a stand-alone method their use can also be integrated with other methods of inquiry. Figure 10 provides a summary overview of the potentially useful relationships between Decision Trees and four other methods mentioned in this paper: Ethnographic Decision Tree Modelling, Hierarchical Card Sorting, network analysis and QCA.

Figure 10: Relationships between data and methods of analysis



Evaluation Applications

Up to now Decision Trees appear to have had little or no use as evaluation tools, either as particular kinds of representations and/or as methods that generate those representations (using computerised algorithms, participatory processes or ethnographic skills). In this final section of the paper I will outline some of the possible uses, which readers might want to experiment with. These fall into three broad categories:

- The elicitation of Theories of Change that might then be evaluated by various means
- The analysis of data that becomes available in the process of project implementation
- The meta-analysis of data from multiple evaluations

The examples discussed below involve data about numbers of cases that range from a dozen up to a thousand or more. The larger sets are amendable to analysis using statistical analyses and the smaller sets are amenable to small-N methods like QCA. They are all amenable to Decision Tree algorithms.

Eliciting tacit and multiple Theories of Change

- **Analysis of data from multiple project locations and implementing bodies**

The DFID-funded Madhya Pradesh Rural Livelihoods Project is a good example of a common project structure that presents problems for use of a single Theory of Change as an evaluation tool. The project has been implemented in 9 tribal districts covering 2901 villages, and has reached an estimated 670,000 households. The overall goal is to "address the livelihood needs of the poorest people in Madhya Pradesh, living in tribal areas" primarily by transferring project funds directly to the village assemblies (Gram Sabhas) who make their own choices about appropriate development activities, albeit within some agreed boundaries. These include livestock and crop support, soil and water conservation, improved management of key natural resources, promotion of rural enterprise, and financial services (including savings, credit, insurance and money transfers). In each state, and within some states, there are different local NGOs providing capacity building support to the Gram Sabhas and their surrounding communities. There are in effect multiple local Theories of Change being pursued through the use of DFID funding.

While there is a LogFrame for the project as a whole the indicators therein do not do justice to the diversity that is present in the project. There are performance measures for the delivery of outputs and the achievement of expected outcomes and impacts. But these are all in the form aggregate measures – total numbers, percentages and averages. Variations from one village to another are in effect being treated as statistical noise. The focus on aggregate measures denies the agency of the very people who the project is targeting.

A Decision Tree analysis could cope with this scale and diversity of contexts, interventions and outcomes, and help generate some generalised conclusions about the configurations of contexts and interventions that are most often successful and unsuccessful. The cases under examination could be the Gram Sabhas, and these could include both those receiving grants from the project and others who may or not be receiving funding from other sources. Alternately it would be possible to do a two stage analysis with districts being the cases of interest in the first stage, if there were good reasons for expecting performance differences between districts and a need to learn about these[37].

- **Testing the often tacit theory built into grant giving mechanisms.**

Donor NGOs such as Comic Relief or the Big Lottery Fund often have quite detailed procedures for screening and then selecting development projects for funding. Some of the selection criteria and processes used are about strategic direction, about what will and will not be funded. Others embody theories about "what will work", sometimes explicitly but often implicitly. These views can concern the nature of the organisation involved and the details of the project design. Rarely, at least in my experience, is the predictive value of these views tested in any systematic way.

This is a setting where Decision Tree analysis should be both possible and relevant. Training cases would be the screened and approved proposals. Their attributes could include the type of organisation implementing the project, the kind of project interventions involved, the kinds of beneficiaries and aspects of the local and national context. Associated outcome measures could be collated after projects have been implemented, using project progress reports and evaluations. The results of a Decision Tree analysis are likely to identify multiple configurations of factors that account for good and not so good performance.

The test cases would be the next tranche of proposals. Does the Decision Tree model developed using the training cases accurately predict the relative success of the new set of projects that are funded? Tested

[37] For example, there are different NGOs working with Gram Sabhas in each district, whose different working methods could have consequences. There are also district level project management committees who may differ in their capacities and priorities.

Decision Tree models could subsequently be used to assess non-funded proposals, to identify possible missed opportunities.

- **Identifying theor<u>ies</u> of change, when there are multiple stakeholders**

In many aid agencies a portfolio of projects may be developed over an extended period of time as a result of decisions made by different people. In these circumstances it is especially likely that there will be multiple theories of change, held by different stakeholders with different relationships to the projects in the portfolio. It is also unlikely that there will be a data set available with comparable project attributes, as they might be in the case of grant giving mechanisms. However data could be collected via e-surveys of stakeholders about their *perceptions* of the kinds of causal processes at work in the various projects. The cases in this situation would be survey respondents rather than projects or grant recipients. The values given to attributes and outcomes would be derived from survey respondents' responses to multiple choice questions. These could cover the kinds of outcomes expected (or not), the kinds of project interventions expected to be most effective (or not) and aspects of the context in which the projects were located which might be conducive or constraining.

A Decision Tree produced as a result of an analysis of this kind of data would capture the aggregate views of all the stakeholders, of what conditions were most likely to be seen to be associated with what outcomes. Because of the diversity of stakeholders and projects a Decision Tree is likely to show multiple configurations i.e. *theories* of change rather than one theory of change, but some with wider support than others. The contents of these configurations, in the form of association rules, should be testable by subsequently searching out for evidence "on the ground" that confirms whether such associations exist or not. This of course would need to be complimented by a search for plausible or tested explanations of the causal mechanisms underlying any associations that were found.

The analysis of project generated data

- **Targeting of poverty alleviation assistance, using data on attributes of poor households to classify them as in or out of the target group**

The following example is an opportunistic analysis of the kind of survey data that could be collected as part of a baseline data collection exercise. In 2006 a poverty survey was carried out in Ha Tinh province, covering 596 households in five communes. The survey instrument, called a Basic Necessities Survey, generated household poverty scores based on possession or absence of various items and access to various services, which were weighted by respondents' collective views of their importance (Davies 2007).

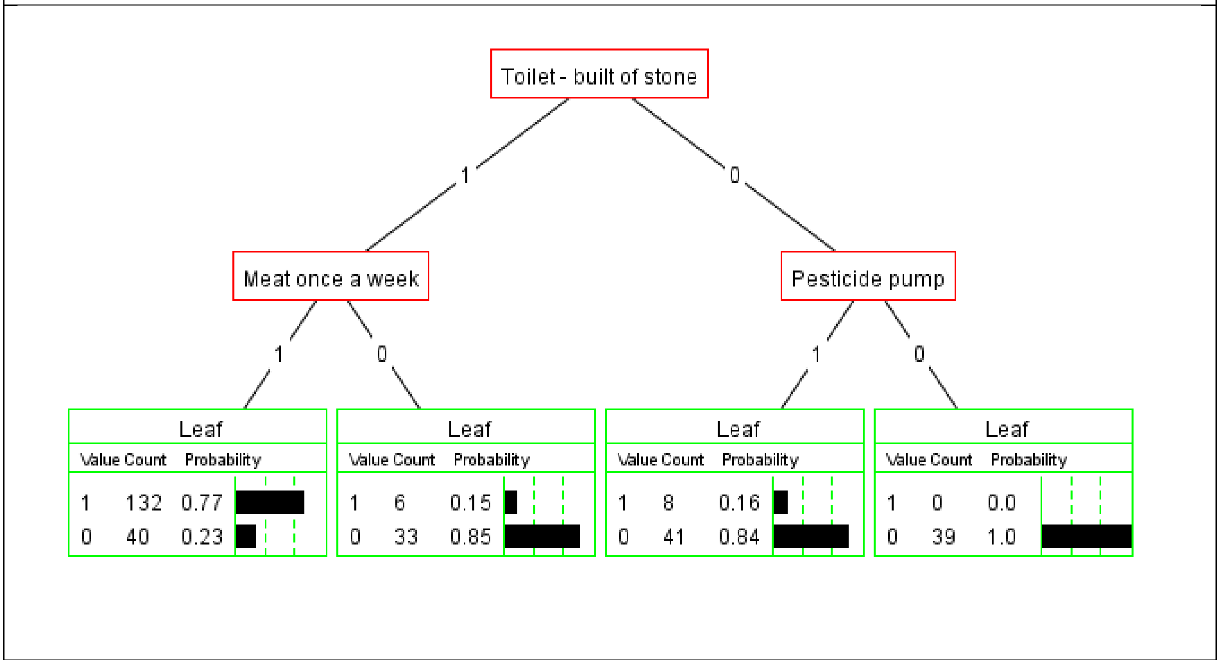
Half of the survey data set was recently used to generate the Decision Tree shown in Figure 11[38]. If it was used as a beneficiary targeting tool, this Decision Tree analysis can be read as saying non-poor households will have a "toilet built of stone" and "eat meat once a week", and all the rest will be poor households. However, as we can see by reading the leaves of the Decision Tree, doing so will involve some errors: 23% of the non-poor households will in fact be poor and 11% of the poor households will in fact be non-poor. These sort of errors can be described and measured in terms of sensitivity and specificity[39].

[38] Using the free dTRee software available at http://aispace.org/dTree/

[39] (from Wikipedia) Sensitivity measures the proportion of actual positives which are correctly identified as such. Specificity measures the proportion of negatives which are correctly identified. These two measures are related to the concepts of Type I and Type II errors.

The Decision Tree was then tested against the other half of the data set, to see how well it correctly identified the poverty status of those households. Its overall accuracy was 82%, which is reasonable but maybe not sufficient for targeting purposes. Increasing the depth of the tree is one potential means of improving accuracy, but in this case it did not do so until there were at least 8 decision levels (versus the two levels below), when the accuracy reached 86%. It is possible that accuracy could be improved further if changes were made in the contents of the household survey instrument e.g. by including more of different items. However there may be limits on how much accuracy can be improved because the subjects of this survey do have individual agency, they some freedom to decide what to buy, even within their very limited incomes[40]. That agency is likely to be responsible for at least some of the residual prediction error. Targeting strategies may always be imperfect.

Figure 11: A Decision Tree generated from household poverty data collected from Ha Tinh province of Vietnam in 2006



Link = 0 = case did not have the attribute Link = 1 = case did have the attribute	Cell = 0 = Household was poor (low BNS score) Cell = 1 = Household was not poor (high BNS score)

- **Content analysis of collections of stories of change**

Stories of change, as reported by project participants, are often collected by aid agencies through various means including the use of the Most Significant Change (MSC) technique, or more recently, through use of the Sensemaker© package. With the latter, the story tellers tag their own stories, using pre-set options. With the former, facilitators can either get participants to categorise stories or they do coding of the stories themselves. While the analysis of the relationships between tagged stories can be theory led and aided by software packages like Nvivo, there is an inherent problem with the scale of the task. With some SenseMaker applications[41] can be 20 or more coding choices, creating a huge combinatorial space, in which they may be multiple potentially meaningful associations between story attributes. The challenge is how to

[40] Agency may also be visible in the form of multiple branches (i.e. configurations) of possession of items, rather than one single branch. In the most extreme case, a separate branch for each respondent.
[41] See Global Giving Story Telling Tools at http://www.globalgiving.org/story-tools/

find them. In addition the number of stories being collected may be very large. This is an area where Decision Tree algorithms can be useful. For example, as a means of finding what story attributes are associated with what kinds of outcomes. GlobalGiving, an American NGO, is now exploring their use[42].

- **Analysis of website usage patterns**

Complex and extensive web sites are now a commonplace feature of many aid agencies, both those managed by governments and NGOs. All websites accumulate detailed datasets each day by automatically recording the actions of each visitor, including when they arrived, on what webpage, how long they were there, what page they went to next, what docs they downloaded, etc., until they leave the website. Understanding the routes different users to reach and use different website contents is potentially relevant to efforts to improve the design of websites, both to direct traffic to specific sections or documents and to increase the length of stay on the website. Decision Tree software can analyse visitor logs and come up with best fitting association rules that will identify what route a visitor is most likely to take to visit a given page, or how long they will stay on the website (Suneetha and Krishnamoorthi 2011; Pabarskaite 2003).

The meta-analysis of data from multiple evaluations

- **Carrying out systematic reviews of evaluations or impact assessments**

Systematic reviews have been used in the fields of medicine for decades to identify and synthesis the findings of studies on a given topic. They are now receiving attention by development aid agencies, including by DFID and AusAID who have funded 3ie to carry out or commission a large number of systematic reviews in recent years. While there are statistical tools for systematically meta-analysing quantitative data from experimental trials there are no such tools for analysing the results of studies and evaluations where they are stated in qualitative terms i.e. in text descriptions[43]. Yet this kind of data is much more widely available. There is also a need for such systematic reviews to generate results in reasonably nuanced form, beyond binary statements about whether an intervention works or does not work, or in terms of a few "treatment-response" rates. A recent issue of the Journal of Development Effectiveness focusing on systematic reviews has included discussion of possible alternatives, but none seem to offer any replicable systematic process.

There have been some exploratory applications of QCA, which hold out the potential to identify from evaluation reports or research studies the multiple configurations of different conditions and circumstances that can generated the expected outcomes. Recently Sager and Andereggen (2012) used QCA to carry out a systematic review of 17 transport policy evaluations in Switzerland. It should be possible to apply a Decision Tree analysis of the same data set (e.g. for triangulation purposes) or to other similar data sets as an alternative means of systematic review with the same advantages of being able to recognise multiple causal configurations.

An invitation…

In the section above I have spelled out a range of situations where the use Decision Trees could be useful. However, with the exception of two of these (story analysis and poverty targeting) these proposals are still conjectures. They need testing. After learning about the merits of Decision Tree models and how Decision Tree algorithms work I hope some readers will now be encouraged to do so.

[42] See Using BigML to dissect trends in 43,388 stories http://chewychunks.wordpress.com/2012/08/14/using-bigml-to-dissect-trends-in-43388-stories/

[43] The recent issue of

and a reminder.

Decision Trees:

- Can use the most widely available form of data (nominal)
- Do not need to make any assumptions need to be made about data distributions and relationships
- Can be used with small and large data sets
- Can present results in a form that is readable by ordinary mortals
- Can include multiple configurations of causes
- Can differentiate causal roles (subject to proviso on associations)
- Can produce results that are testable
- Their performance is evaluable.

Postscript 1: Schneider and Grofman (2006) have explored alternate ways of representing QCA results in more user friendly forms, including dendrograms (i.e. decision trees). They see dendrograms as having the potential to display temporal or causal pathways, similar to what was suggested by Gladwin's guidance on the development of Ethnographic Decision Tree Models. However Decision Tree models as developed by data mining algorithms do not seek to capture this dimension. Efforts to do so seem likely to risk reducing their performance as classifiers and predictors. Schneider and Grofman are in agreement with this paper in seeing Decision Trees as good means of representing equifinality.

Postscript 2: Fiss (2012) has presented data on the attributes of 13 high and low performing private sector organisations. Using QCA he was able to identify two rules that correctly classify the high performers. When the same data set was analysed using dTree and Rapid Miner these programs identified three rules, two of which had been identified by Fiss. Assessed in terms of their relative simplicity (Occam's Razor) Fiss's QCA solution was the better. This result is in contrast to the other example of QCA versus Decision Tree results given on page 18-19 above (re Fischer's results), where the Decision Tree produced fewer fitting rules. Of course, simplicity is not the only relevant criteria. Inquiries need to be made in both cases as to which have the most plausible causal mechanisms underlying the association rules.

References

Backhouse, R. E., and M. S. Morgan. 2000. 'Introduction: Is Data Mining a Methodological Problem?' *Journal of Economic Methodology* 7 (2): 171–181.

CARE. 2012. *Defining Theories of Change*. http://www.careinternational.org.uk/research-centre/conflict-and-peacebuilding/155-peacbuilding-with-impact-defining-theories-of-change.

Chen, Xiaojun, Graham Williams, and Xiaofei Xu. *A Survey of Open Source Data Mining Systems*.

Davies, Rick. 1996. 'Hierarchical Card Sorting: A Tool for Qualitative Research'. *Monitoring and Evaluation NEWS*. http://www.mande.co.uk/docs/hierarch.htm.

———. 2007. 'The 2006 Basic Necessities Survey (BNS) in Can Loc District, Ha Tinh Province, Vietnam'. Pro Poor Centre, Vietnam. http://mande.co.uk/blog/wp-content/uploads/2012/11/The-2006-Basic-Necessities-Survey-Final-Report-20-July-2007.pdf.

Eguren, Iñigo Retolaza. 2011. 'THEORY OF CHANGE: A thinking and action approach to navigate in the complexity of social change processes'. HIVOS.

http://www.democraticdialoguenetwork.org/file.pl?files_id=1811;folder=attachment;name=Theory_of_Change.pdf.

Fischer, Manuel. 2011. 'Social Network Analysis and Qualitative Comparative Analysis: Their Mutual Benefit for the Explanation of Policy Network Structures'. Methodological Innovations Online. http://unige.ch/ses/spo/Membres/Enseignants/Fischer/Publications/7FEEDd01-1.pdf.

Fiss, Peer C. 2009. "Practical Issues in QCA". Academy of Management. http://www-bcf.usc.edu/~fiss/QCA_PDW_2009_Fiss_Practical_Issues.pdf.

Fiss, Peer C. 2012. 'A Set-Theoretic Approach to Organizational Configurations'. *SSRN eLibrary*. Accessed November 26. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1004664.

Funnell, Sue C., and Patricia J. Rogers. 2011. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. John Wiley & Sons.

Gladwin, Christina H. 1989. *Ethnographic Decision Tree Modeling*. Sage.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton University Press. http://www.amazon.co.uk/Tale-Two-Cultures-Qualitative-Quantitative/dp/0691149712/ref=sr_1_1?ie=UTF8&qid=1353850106&sr=8-1.

Harloff, J., and A. P. M. Coxon. 2005. *How To Sort*. http://methodofsorting.com/HowToSort1-1_english.pdf.

James, Cathy. 2011. 'THEORY OF CHANGE REVIEW A report commissioned by Comic Relief'. Comic Relief. http://mande.co.uk/blog/wp-content/uploads/2012/03/2012-Comic-Relief-Theory-of-Change-Review-FINAL.pdf.

Krook, M. L. 2010. 'Women's Representation in Parliament: A Qualitative Comparative Analysis'. *Political Studies* 58 (5): 886–908.

Mayne, John. 2012. 'Contribution Analysis: Coming of Age?' *Evaluation* 18 (3) (July 1): 270–280. doi:10.1177/1356389012451663.

Moore, T., C. Jesse, and R. Kittler. 2001. 'An Overview and Evaluation of Decision Tree Methodology'. In *American Statistical Association Quality and Productivity Conference Papers. University of Texas, Austin*. http://www.amstat-online.org/sections/qp/qpr/QPRC2001/contributed/Moore.pdf.

Pabarskaite, Zidrina. 2003. 'Decision Trees for Web Log Mining'. *Intell. Data Anal.* 7 (2) (April): 141–154.

Patton, Michael Quinn. 2012. 'A Utilization-focused Approach to Contribution Analysis'. *Evaluation* 18 (3) (July 1): 364–377. doi:10.1177/1356389012449523.

Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation*. SAGE.

Pritchett, Lant, Salimah Samji, and Jeffrey Hammer. 2012. 'It"s All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space'. http://giving-evidence.com/2012/08/09/worms/.

Ragin, Charles C. 1987. *The comparative method : moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Rokach, Lior, and Oded Z. Maimon. 2008. *Data Mining with Decision Trees: Theroy and Applications*. World Scientific.

Ryan, G. W., and H. R. Bernard. 2006. 'Testing an Ethnographic Decision Tree Model on a National Sample: Recycling Beverage Cans'. *Human Organization* 65 (1): 103–114.

Sachdeva, K., M. Hanmandlu, and A. Kumar. 2012. 'Real Life Applications of Fuzzy Decision Tree'. *International Journal of Computer Applications* 42 (10): 24–28.

Sager, Fritz, and Céline Andereggen. 2012. 'Dealing With Complex Causality in Realist Synthesis The Promise of Qualitative Comparative Analysis'. *American Journal of Evaluation* 33 (1) (March 1): 60–78. doi:10.1177/1098214011411574.

Spradley, James P. 1979. *The Ethnographic Interview*. Wadsworth Publishing Co Inc.

Stein, Danielle, and Craig Valters. 2012. 'Understanding Theory in Change in International Development'. The Asia Foundation. http://www2.lse.ac.uk/internationalDevelopment/research/JSRP/downloads/JSRP1.SteinValters.pdf.

Stern, Leon, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, and Barbara E. Befani. 2012. 'BROADENING THE RANGE OF DESIGNS AND METHODS FOR IMPACT EVALUATIONS Report of a study commissioned by the Department for International Development'. DFID. http://www.dfid.gov.uk/r4d/pdf/outputs/misc_infocomm/DFIDWorkingPaper38.pdf.

Suneetha, K. R., and R. Krishnamoorthi. 2011. 'Classification of Web Log Data to Identify Interested Users Using Decision Trees'. *International Journal of Advanced Computer Science and Applications (IJACSA)* 2 (12). http://ubicc.org/files/pdf/Classn_439.pdf.

Vogel, Isabel. 2012. 'Review of the use of "Theory of Change" in international development. Review Report'. DFID. http://www.dfid.gov.uk/r4d/pdf/outputs/mis_spc/DFID_ToC_Review_VogelV7.pdf.

White, Howard. 2011. 'An Introduction to the Use of Randomized Control Trials to Evaluate Development Interventions'. 3ie. http://www.3ieimpact.org/media/filer/2012/05/07/Working_Paper_9.pdf.

———. 2012. 'Exercising credibility: why a theory of change matters'. *Blog-3ie:International Initiative for Impact Evaluation | Evaluating Impact, Informing Policy, Improving Lives*. http://www.3ieimpact.org/en/blog/2012/08/28/exercising-credibility-why-theory-change-matters/.