FUZZY SET QUALITATIVE COMPARATIVE ANALYSIS

THOMAS ELLIOTT

1. Introduction

fsQCA is, fundamentally, an analysis of set relations.

Sets are groups of things. In formal set theory, sets are usually composed of numbers, or other sets. But we can also think of sets of things. For example, the set of all mammals. Or the set of countries with populations over 10 million.

Sets can be subsets of larger sets. For example, the set of grad students is a subset of the set of university students. The set of cats is a subset of the set of mammals. Sometimes, these subsets are definitional (as in the previous examples), but sometimes they can be causal. For example, the set of students who studied constitute a subset of students who did well on the test. The set of zip codes with high property values is a subset of zip codes where lots of people live.

In fsQCA, we transform our variables into sets. We then analyze what combination of causal sets constitute a subset of the outcome set.

fsQCA differs from regression analysis in the way it focuses on problems and comes up with solutions. In regression analysis, the goal is to discover the effect a variable as on some outcome. Regression gives us the magnitude and direction of effect of a variable, net of other variables included in the model. However, in fsQCA, the focus is on what conditions lead to a given outcome.

There are many advantages fsQCA has over traditional correlational analysis like regression.

- Relationships are asymmetrical
- Equifinality
- Causal complexity

1.1. **Asymmetry.**

Set relations are asymmetrical. Consider the following made up data:

Table 1. Sample Data

	Conservative	Liberal
Sociologist	2	25
Not Sociologist	30	26

If you calculated a correlation coefficient from the contingency table above, the high n in the lower right cell would diminish the correlation's magnitude. But a set-theoretic claim that sociologists are politically liberal is not diminished by the large number of liberals who are not sociologists.

1.2. **Equifinality.**

Equifinality means that there are multiple paths or solutions to the same outcome. For example, there can be multiple ways to get into college. One pathway involves studying hard and getting good grades. Another way is being rich and having connections.

Date: January 15, 2013.

1.3. **Causal Complexity.** In regression analysis, you are finding the net effects of variables - what is the independent effect of variable A, net of all other variables?

fsQCA finds combinations of causal measures that lead to the outcome. In fsQCA, it's not about independent effects, but combined effects. For example,

2. fsQCA

Qualitative comparative analysis entails the analysis of necessary and sufficient conditions to produce the some outcome. Necessary conditions are conditions that are required to produce the outcome. All cases that exhibit the outcome also exhibit a necessary condition. Though, necessary conditions may not be enough by itself. It set notation, you can think of the outcome set as a subset of the necessary condition set.

Sufficient conditions are conditions that always lead to the outcome. So cases that exhibit the sufficient condition will also exhibit the outcome. Sufficient conditions may not be the only conditions that lead to the outcome, however. In set notation, you can think of the sufficient condition set as a subset of the outcome set.

To perform fsQCA, we just must convert our variables into sets. Sets can be either crisp or fuzzy.

Crisp sets denote sets in which membership is either on or off. Consider the set of people who have a PhD - you either have one or you don't. The boundary defining the set is crisp.

Fuzzy sets are sets in which membership can be expressed in degrees. Consider the set of people who are rich. We might express a degree of membership: some people are not rich, some people are slightly rich, and some people are über rich.

2.1. **Calibration.** Both crisp and fuzzy sets require calibration - deciding how you will define membership in the set. For some things (like the set of people with PhDs) this is pretty straight forward. But for other sets, it requires thought and theory.

What constitutes über rich? When is an economy considered developed? What is a high GDP?

Cut off points for membership should be grounded in both theory and the cases you are studying.

Fuzzy sets have a special point in between full membership and full non-membership: the crossover point.

A membership score of 0.5 denotes the cases with the maximum ambiguity about their membership in the set. All three points should be theoretically and empirically grounded.

2.2. **Negated Sets.** In fsQCA, we have negated sets, or the absence of a set. We denote negated sets by $\sim$A, or sometimes, the lowercase of the set name. We can calculate the membership of a case in a negated set by taking one minus the membership score.

For example, say you have a case with a membership score of 0.7 in the set of people who are rich. We can negate the set, so that it is the set of people who are not rich. Our case would have a membership score in the negated set of $1 - 0.7 = 0.3$

3. An Example

Consider this example with two fuzzy causal sets and one fuzzy outcome set:

Causal sets:

- People who are Rich
- People who are Educated

Outcome set:

- People who are Liberal

Table 2. Example Data

	Rich	Educated	Liberal
David	0.2	0.3	0.3
Kathy	0.3	0.7	0.8
Mitch	0.9	0.4	0.1
Eileen	0.8	0.9	0.7

3.1. **Truth Table.** Analysis of fuzzy set data revolves around the truth table. The truth table consists of all possible combinations of causal sets, one row for each combination. This means that if you have k causal sets, your truth table will have 2^k rows. We use the truth table to assess which combinations of causal conditions lead to the outcome

Table 3. Truth Table

Rich	Educated
True	True
True	False
False	True
False	False

The first thing we need to do is assign cases to truth table rows. For crisp sets, assigning cases to truth table rows is straightforward. For fuzzy sets, though, it is more complicated. What we need to do is assign a membership score for the causal combination as a whole. We set this membership score to the lowest membership score of the individual sets in the combination.

Consider Kathy and the causal combination Rich*Educated. Kathy has a Rich membership of 0.3 and an Educated membership of 0.7. Kathy's Rich*Educated membership is the lowest membership score of the individual sets, which is 0.3.

Now consider Kathy and ∼Rich*Educated. Remember, ∼Rich is the negated set for Rich. So Kathy's ∼Rich membership score is $1-0.3=0.7$, which is also her membership score for Educated. So Kathy's membership score for ∼Rich*Educated is 0.7.

Cases are assigned to the combinations in which they have membership scores greater than 0.5 (see table 4).

In our example, we have one case per row. Sometimes, though, some rows will have no cases. These are designated as remainders. If you have a lot of cases, you may also want to designate rows as remainders that have only one or two cases. These decisions should be guided by your knowledge of your cases.

TABLE 4. Membership Scores

	Rich* Educated	∼Rich* Educated	Rich* ∼Educated	∼Rich* ∼Educated
David	0.2	0.3	0.2	**0.7**
Kathy	0.2	**0.7**	0.3	0.3
Mitch	0.4	0.1	**0.6**	0.1
Eileen	**0.8**	0.2	0.1	0.1

TABLE 5. Truth Table

Rich	Educated	Cases
True	True	1
True	False	1
False	True	1
False	False	1

3.2. **Consistency.** In fsQCA, consistency represents the extent to which a causal combination leads to an outcome. Consistency ranges from 0 to 1. With crisp sets, you can think of it as the proportion of cases with a given causal combination that are also in the outcome set. For fuzzy sets, it's a similar concept, but calculated differently.

$$Consistency(X < Y) = \frac{\sum \min(X, Y)}{\sum X}$$

Where X is the membership score in causal combination and Y is the membership score in the outcome set.

TABLE 6. Calculating Consistency

	Rich* Educated (X)	Liberal (Y)	Min(X,Y)
David	0.2	0.3	0.2
Kathy	0.2	0.8	0.2
Mitch	0.4	0.1	0.1
Eileen	0.8	0.7	0.7

$$\frac{0.2 + 0.2 + 0.1 + 0.7}{0.2 + 0.2 + 0.4 + 0.8} = 0.75$$

Once we've calculated consistency scores for all causal combinations, we then decide which combinations we will include in the final solution. We do this by picking a cut-off values for consistency scores, those rows with high enough scores we keep for our solution. Rows with high consistency indicate combinations that almost always leads to the outcome.

The exact cutoff should be driven by your data. You should look at the truth table and see if there are any natural cutoff points indicated by large gaps in consistency moving down the table. 0.8 is a good starting point to look at, but you can try different cutoff points to see how it affects your results. The higher your cutoff point, the higher your final consistency will be, but the lower your coverage will be (we'll discuss coverage next).

TABLE 7. Truth Table, sorted by consistency

Rich	Educated	Cases	Consistency	Include
False	True	1	1.00	1
True	True	1	0.75	0
False	False	1	0.67	0
True	False	1	0.58	0

So this gives us a solution of ∼Rich*Educated. Remember, we are investigating what causal factors lead to being liberal. So here, not being rich combined with being educated leads to being liberal.

3.3. **Coverage.** Once we've chosen which rows have high consistency, we can calculate a second statistic: coverage. Coverage represents how many cases with the outcome are represented by a particular causal condition. Since we are assuming that the causal conditions lead to the outcome, it only makes sense to calculate coverage for rows that have high consistency. Rows with low consistency violate our assumption that the causal condition leads to the outcome.

Calculating coverage is similar to calculating consistency:

$$Coverage = \frac{\sum \min(X, Y)}{\sum Y}$$

TABLE 8. Calculating Coverage

	∼Rich* Educated (X)	Liberal (Y)	Min(X,Y)
David	0.3	0.3	0.3
Kathy	0.7	0.8	0.7
Mitch	0.1	0.1	0.1
Eileen	0.2	0.7	0.2

$$\frac{0.2 + 0.7 + 0.1 + 0.2}{0.3 + 0.8 + 0.1 + 0.7} = 0.63$$

So our final solution is ∼Rich*Educated, which has a consistency of 1.00 and a coverage of 0.63.

We could have decided to include the top two rows of the truth table sorted by consistency. So that our solution includes both ∼Rich*Educated and Rich*Educated as leading to being Liberal. We can write this as:

$$\sim \text{Rich} * \text{Educated} + \text{Rich} * \text{Educated} = \text{Liberal}$$

Here, the $*$ indicates logical AND and $+$ indicates logical OR. Notice, though, that our solution has both ∼Rich and Rich — you can be educated and not rich or you can be educated and rich. We can simplify this, because it doesn't matter whether you are rich or not, as long as you are educated.

$$\text{Educated} = \text{Liberal}$$

This simplification process is very useful, especially when you have more causal measures.

Let's calculate our consistency and coverage for this simplified solution:

TABLE 9. Calculating Consistency & Coverage

	Educated (X)	Liberal (Y)	Min(X,Y)
David	0.3	0.3	0.3
Kathy	0.7	0.8	0.7
Mitch	0.4	0.1	0.1
Eileen	0.9	0.7	0.7

$$Consistency = \frac{0.3 + 0.7 + 0.1 + 0.7}{0.3 + 0.7 + 0.4 + 0.9} = 0.783$$

$$Coverage = \frac{0.3 + 0.7 + 0.1 + 0.7}{0.3 + 0.8 + 0.1 + 0.7} = 0.947$$

Notice that our consistency has decreased, but our coverage has increased. This is a normal trade off — the lower your consistency, the higher your coverage. The goal is to find a good balance, in which the solution is empirically and theoretically compelling, and your consistency and coverage are in ranges that validate your solution. If you have a super high consistency, but your coverage is super low, then your solution isn't that compelling because it doesn't describe many cases at all. On the other hand, if you have high coverage, but low consistency, it isn't compelling because your solution doesn't lead to the outcome often enough to make a strong causal argument.