FUZZY SET QUALITATIVE COMPARATIVE ANALYSIS PART 2

THOMAS ELLIOTT

These instructions build on a previous handout, which give an introduction to the basics of set theory and fsQCA. Here, we will use software developed by Charles Ragin to perform fsQCA on a dataset I've collected on newspaper coverage of LGBT social movement organizations.

You can download the fsqca software from here:
`http://www.u.arizona.edu/~cragin/fsQCA/software.shtml`

1. Getting Your Data Ready

To get your data ready for the fsQCA software, you'll need to export it into a comma-separate format. The first row should contain variable names, and the names should not contain spaces or punctuation. Each column of data should have the same data type (so all values in a column are either numeric or strings, but not both). It will be handy to have a column of IDs, either numeric or string, to easily identify individual cases during the analysis, but this is not required.

1.1. **LGBT News Coverage Data.** I will be using a dataset I constructed as examples throughout this handout. The case in this data is the SMO-year, or data for each SMO for each year since it is founded, with the earliest year set at 1969. The outcome measure is the number of newspaper articles the SMO is mentioned in, from the New York Times, the Los Angeles Times, and the Wall Street Journal. I have seven independent variables:

- **Policy Focus** - the policy focus of the SMO. If the SMO has no policy focus, in that it pursues a variety of different policies (think of the Human Rights Campaign or Lambda Legal), they are coded as "all."
- **SMO Identity** - the identity focus of the SMO, based on the LGBT acronym. So an SMO geared primary towards lesbian issues is coded as lesbian. SMOs with no overt identity focus is coded as "all."
- **AIDS Deaths** - the annual number of deaths due to AIDS in the United States.
- **SMO Protest** - whether the SMO engages in protest as a primary tactical tool.
- **SMO Age** - the age of the SMO for the current SMO-year.
- **Policy Score** - this is a cumulative policy score, based on policy gains and setbacks related to LGBT issues. SMOs are assigned a policy score based on their policy focus - so separate scores are calculated for different policy foci, and then assigned to the

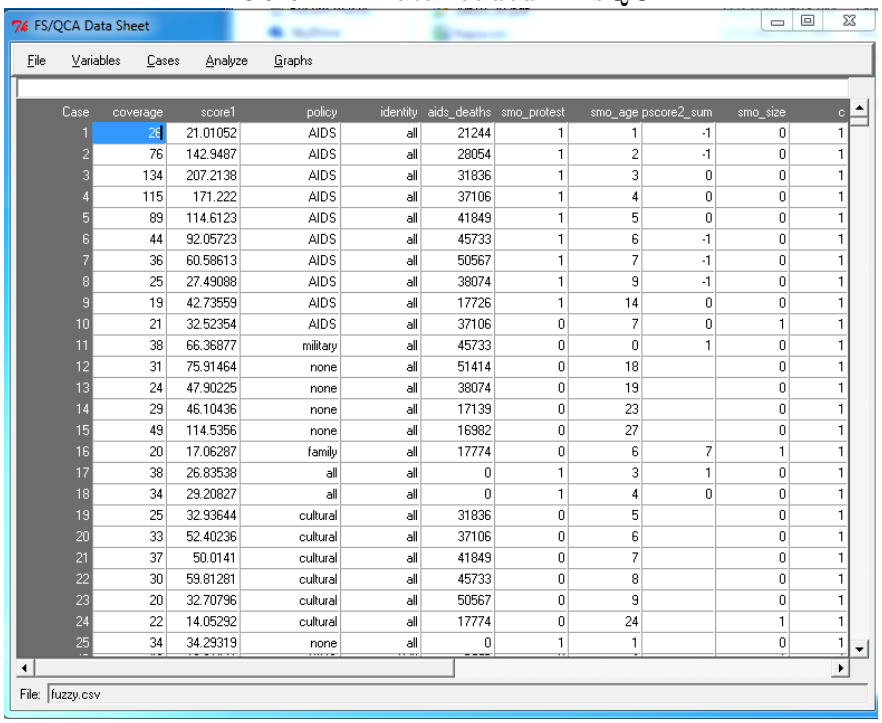
Date: February 25, 2013.

relevant SMO. SMOs coded as "all" on their policy focus measure are given a policy score that is the sum of all individual policy scores.

- **SMO Resources** - a dummy variable indicating an SMO with a large amount of resources.

1.2. **Importing and Calibrating Your Data.** Once you've got your data properly formatted and saved, launch fsQCA and open the data file using the menus. If all goes well, you should see a spreadsheet window containing your data, as in figure 1.

FIGURE 1. Data loaded in fsQCA



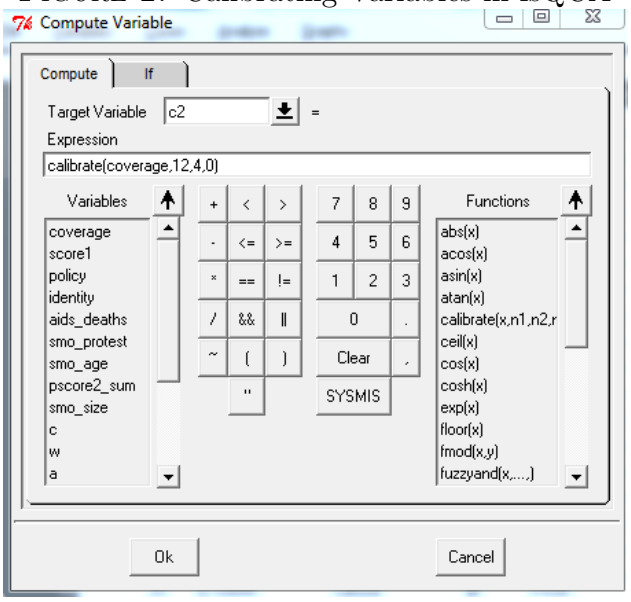
Case	coverage	score1	policy	identity	aids_deaths	smo_protest	smo_age	pscore2_sum	smo_size	c
1	28	21.01052	AIDS	all	21244	1	1	-1	0	1
2	76	142.9487	AIDS	all	28054	1	2	-1	0	1
3	134	207.2138	AIDS	all	31836	1	3	0	0	1
4	115	171.222	AIDS	all	37106	1	4	0	0	1
5	89	114.6123	AIDS	all	41849	1	5	0	0	1
6	44	92.05723	AIDS	all	45733	1	6	-1	0	1
7	36	60.58613	AIDS	all	50567	1	7	-1	0	1
8	25	27.49088	AIDS	all	38074	1	9	-1	0	1
9	19	42.73559	AIDS	all	17726	1	14	0	0	1
10	21	32.52354	AIDS	all	37106	0	7	0	1	1
11	38	66.36877	military	all	45733	0	0	1	0	1
12	31	75.91464	none	all	51414	0	18		0	1
13	24	47.90225	none	all	38074	0	19		0	1
14	29	46.10436	none	all	17139	0	23		0	1
15	49	114.5356	none	all	16982	0	27		0	1
16	20	17.06287	family	all	17774	0	6	7	1	1
17	38	26.83538	all	all	0	1	3	1	0	1
18	34	29.20827	all	all	0	1	4	0	0	1
19	25	32.93644	cultural	all	31836	0	5		0	1
20	33	52.40236	cultural	all	37106	0	6		0	1
21	37	50.0141	cultural	all	41849	0	7		0	1
22	30	59.81281	cultural	all	45733	0	8		0	1
23	20	32.70796	cultural	all	50567	0	9		0	1
24	22	14.05292	cultural	all	17774	0	24		1	1
25	34	34.29319	none	all	0	1	1		0	1

From here we can recode variables, construct new variables, and, most importantly, calibrate[1] variables for fsQCA. To calibrate a variable, go to the Variables menu item and choose compute. This will bring up a window as in figure 2. From this window, you can compute a variety of different new variables, not just fuzzy membership scores.

First, I'm going to calibrate my coverage variable into a membership score for the set of SMO-years with high coverage. In this case, I'm defining high coverage as monthly coverage. I'll name my new variable C in the Target Variable box. Then I'll type `calibrate(coverage,12,4,0)` into the expression box. The calibrate function has four parts. First, the name of the variable you are calibrating — here we are calibrating the coverage variable. Second, the threshold for full membership. Third, the crossover point. Fourth, the threshold for full nonmembership. Clicking okay creates a new variable called C that contains fuzzy membership scores for the set of SMOs with high coverage.

[1]See the previous handout for more thorough discussion of calibration

FIGURE 2. Calibrating Variables in fsQCA



I can also create crisp sets using categorical data. I want to create a crisp set of SMOs who focus on AIDS policies. I again go to Variables and then Compute. I'm going to name the new variable A, and then in the expression box type `policy == ''AIDS''`. Clicking okay will create a new variable that is 1 for rows in which the policy variable was equal to "AIDS," and 0 for all others.

Recoding the variables, I get the following set membership scores:

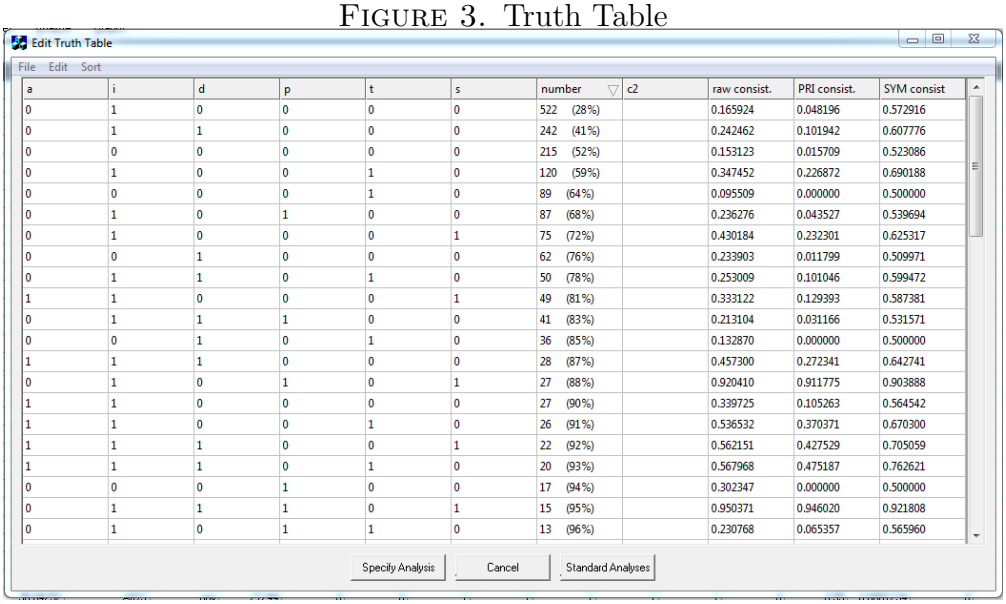
- `C` — the set of SMO-years with high coverage. Fuzzy measure with 0-4-12 thresholds.
- `A` — the set of SMOs that focus on AIDS policy. Crisp.
- `I` — the set of SMOs with an inclusive identity focus (coded "all"). Crisp.
- `D` — the set of years with high death rates due to AIDS. Fuzzy measure with 0-20000-40000 thresholds.
- `S` — the set of SMO-years with high resources. Crisp.
- `T` — the set of SMOs who protest. Crisp.
- `O` — the set of SMO-years that are over 7 years old. Crisp.
- `P` — the set of SMO-years with a high policy score. Fuzzy measure with 3-7-12 for generally focused SMOs, 2-4-7 for SMOs focused on a specific policy.

2. The Truth Table

2.1. **Creating the Truth Table.** To generate a truth table, go to Analyze then fuzzy truth table. The next window asks for the outcome set, which we'll set to C. It will also ask for the causal sets. Initially, we'll set it to A, I, D, T, S, and P. Clicking okay will produce our truth table. Since we have six causal sets, the truth table has $2^6 = 64$ rows. There's a column

for each causal set, with 1 indicating the presence of the set and 0 indicating the absence of the set. Next we have a column containing the number of cases in each row. We have a blank column which we'll use in a minute. Then we have three columns containing different calculations of consistency. We'll focus on the first of these for now. Figure 3 contains the truth table from this analysis.

FIGURE 3. Truth Table



a	i	d	p	t	s	number	c2	raw consist.	PRI consist.	SYM consist
0	1	0	0	0	0	522 (28%)		0.165924	0.048196	0.572916
0	1	1	0	0	0	242 (41%)		0.242462	0.101942	0.607776
0	0	0	0	0	0	215 (52%)		0.153123	0.015709	0.523086
0	1	0	0	1	0	120 (59%)		0.347452	0.226872	0.690188
0	0	0	0	1	0	89 (64%)		0.095509	0.000000	0.500000
0	1	0	1	0	0	87 (68%)		0.236276	0.043527	0.539694
0	1	0	0	0	1	75 (72%)		0.430184	0.232301	0.625317
0	0	1	0	0	0	62 (76%)		0.233903	0.011799	0.509971
0	1	1	0	1	0	50 (78%)		0.253009	0.101046	0.599472
1	1	0	0	0	1	49 (81%)		0.333122	0.129393	0.587381
0	1	1	1	0	0	41 (83%)		0.213104	0.031166	0.531571
0	0	1	0	1	0	36 (85%)		0.132870	0.000000	0.500000
1	1	1	0	0	0	28 (87%)		0.457300	0.272341	0.642741
0	1	0	1	0	1	27 (88%)		0.920410	0.911775	0.903888
1	1	0	0	0	0	27 (90%)		0.339725	0.105263	0.564542
1	1	0	0	1	0	26 (91%)		0.536532	0.370371	0.670300
1	1	1	0	0	1	22 (92%)		0.562151	0.427529	0.705059
1	1	1	0	1	0	20 (93%)		0.567968	0.475187	0.762621
0	0	0	1	0	0	17 (94%)		0.302347	0.000000	0.500000
0	1	1	1	0	1	15 (95%)		0.950371	0.946020	0.921808
0	1	0	1	1	0	13 (96%)		0.230768	0.065357	0.565960

2.2. **Delete Empty Rows.** The first thing we want to do is drop any rows that don't have enough cases to include in the analysis. Typically, we just drop rows with no cases, though if you have enough data, you could also drop rows that only have one or two cases. The truth table should already be sorted by the number of cases in each row. Simple click on a cell in the first row with no cases, then go to Edit and then delete this row to the end. This will delete the current row and all rows below it. These rows are designated remainders, since they don't contain any data, and can be used to help simplify solutions, which we'll discuss later.

2.3. **Code the Outcome.** Now we need to choose a consistency cut off for when a causal path leads to the outcome. First, sort by the raw consistency — click somewhere in the raw consistency column, then click Sort and descending. Now the table should be sorted by consistency in descending order. We want to look for natural breaks in the consistency — look for large gaps in consistency as you go down. In this truth table, there is a gap between 0.92 and 0.62, so this seems like a natural break point. No we need to fill in the empty column with 1s for rows that are consistent with the outcome, and 0s for those that are not. We can do this manually, or go to Edit then Delete and Code. This let's us automatically delete any rows that have less than a given number of cases, and automatically code the outcome based on a given consistency. We can put in 0.8 to automatically code everything above a consistency of 0.8 as having the outcome.

3. The Analysis

Once we've coded the outcome variable in our truth table, we can perform our analysis. Click on the Standard Analysis button. The software then minimizes the rows of the truth table using Boolean logic. Sometimes, as in our case, the software will ask for manual input in choosing prime implicants.

3.1. **Prime Implicants.** The software will automatically minimize causal conditions from the truth table to create simpler solutions. The parsimonious solution also uses the rows designed as remainders (more on that later). Often, the software will find simplified solutions that are redundant. For example, let's say you have a three causal sets–A, B, and C–to explain an outcome set, O. And your final solution contains the following rows in the truth table (called primitive expressions):

$$O = A * b * C + a * B * c + A * B * c + A * B * C$$

These four terms can be minimized in the following way:

$A * B * C$ combines with $A * b * C$ to produce $A * C$
$A * B * C$ combines with $A * B * c$ to produce $A * B$
$A * B * c$ combines with $a * B * c$ to produce $B * c$

$$O = A * C + A * B + B * c$$

These three terms are called prime implicants. However, notice that $A * B$ implies $A * B * C$ and $A * B * c$, but these two are also implied by $A * C + B * c$. $A * B$ is redundant, then — we don't need it in the solution to cover all the rows of the truth table. When this happens, the software will ask us to choose among the redundant prime implicants to include in the solution. We must choose the minimum number of implicants required to cover all the primitive expressions, but we can include additional implicants if we think they are important theoretically or empirically.

Figure 4. Prime Implicant Window

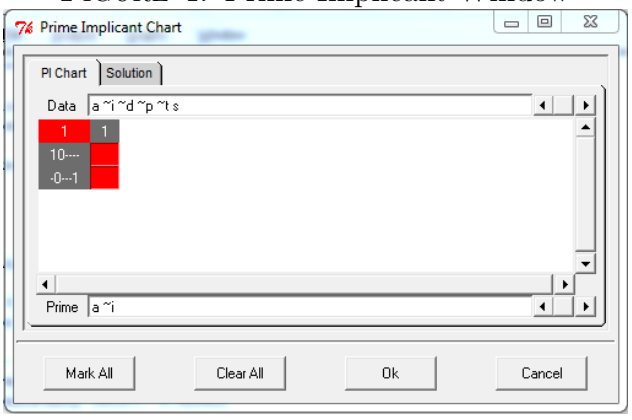


Figure 4 shows the window that will pop up if you need to choose among redundant prime implicants. The data field shows the primitive expression that is covered by redundant

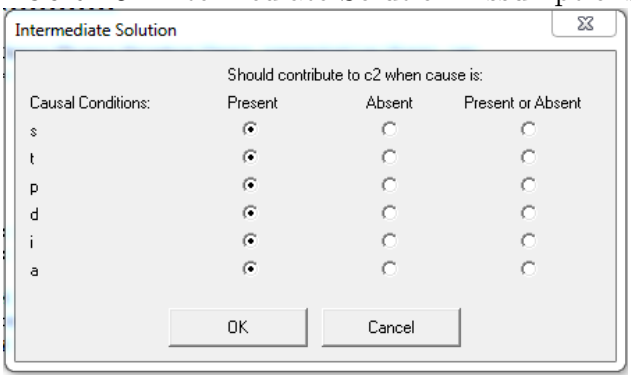
implicants. The rows of the implicant table correspond to redundant prime implicants. Clicking in the cell will show the implicant in the bottom field. Clicking on the row header will mark that implicant as being included in the solution. The numbers along the top of the columns indicate how many implicants you must choose before the software can proceed.

The first row corresponds to the prime implicant $A*i$, and the second corresponds to $i*S$. I will choose $i*S$ because that is a more theoretically interesting solution than $A*i$. Clicking okay proceeds to the next window.

3.2. **Intermediate Assumptions.** Intermediate solutions use counterfactuals to try to simplify the complex solution without making unjustified assumptions. This requires thinking about how each causal set is expected to contribute to the outcome. For example, if your outcome was revolution, and one of your causal sets was popular unrest, you would expect the presence, not the absence, of popular unrest to contribute to occurrence of revolutions.

When the software is calculating the intermediate solution, a window will pop up asking for filtering assumptions for calculating the intermedia solution. These assumptions should be grounded in your case and in theory. I'll discuss the intermediate solution more later, for now we'll set all the assumptions to present. See figure 5.

FIGURE 5. Intermediate Solution Assumptions



Once we hit okay, the results pop up with the results of our fsQCA. The software computes three different types of analyses: the complex solution, the parsimonious solution, and the intermediate solution.

Note: the software indicates negated sets, or the absence of a set, with $\sim$. In my discussion, I will indicate negated sets with the lowercase letter.

3.3. **The Complex Solution.**

```
--- COMPLEX SOLUTION ---
frequency cutoff: 4.000000
consistency cutoff: 0.920410

                     raw        unique
                  coverage    coverage    consistency
```

```
                 ----------  ----------  ----------
~a*i*p*s         0.148442    0.101810    0.933018
a*~i*~p*~t*s     0.053182    0.053182    0.883736
~a*~d*p*~t*s     0.058208    0.011576    0.920875
solution coverage: 0.213200
solution consistency: 0.919670
```

The complex solution makes no simplifying assumptions. It takes the rows from the truth table coded as one on the outcome, and then applies some boolean simplification to combine rows, but that is it. As a result, if you've included a larger number of causual conditions, you'll get pretty complicated solutions.

So according to the above solution, we have three pathways to high coverage. The first, $a*I*P*S$, indicates that non-AIDS focused organizations with an inclusive identity focus, high policy score, and large resources will gain high coverage. This is fairly consistent at 0.933, and explains the most cases of high coverage of the three pathways, but the coverage is still only 0.148.

The second pathway, $A*i*p*t*S$, indicates that AIDS organizations that do not have an inclusive identity focus, do not have a high policy score, do not engage in protest, and have a large amount of resources are likely to gain high coverage. This pathway is slightly less consistent at 0.884, and has less coverage, 0.0531, than the previous pathway.

The final pathway, $a*d*P*t*S$, indicates that AIDS organizations during a time when AIDS deaths are not high, that have high policy scores, that do not engage in protest, and that have a high number of resources will gain high coverage. This has a consistency of 0.921 and a coverage of 0.0582.

We can combine all three pathways using boolean algebra: $S*(a*P*(I+d*t)+A*i*p*t)$. Obviously this is a complex recipe, and so it may be easier in instances like this to keep the pathways separate. Overall, our solution has a high consistency of 0.920, but a relatively low coverage of only 0.213. However, the solution has a necessary (but not sufficient) condition to high coverage: S. All the pathways require S, so it is necessary to gain high coverage.

3.4. **The Parsimonious Solution.**

```
--- PARSIMONIOUS SOLUTION ---
frequency cutoff: 4.000000
consistency cutoff: 0.920410

            raw       unique
          coverage    coverage   consistency
         ----------  ----------  ----------
p*s      0.164582    0.148654    0.912729
~i*s     0.069061    0.053133    0.850645
solution coverage: 0.217715
solution consistency: 0.905270
```

The parsimonious solution uses any and all remainder rows to help simplify the solution. If a remainder helps create a simpler solution, then it is assumed to have the outcome and

included in the solution. Obviously, this is a strong assumption and so the parsimonious solution should only be used if you are fully satisfied the assumptions made to create the solution are justified.

We can see here that we've reduced down to two pathways, both of which are far simpler than the complex solution. The first, $P * S$, indicates that organizations with high policy scores and large resources will gain high coverage. This has a fairly high consistency, 0.913, and the largest coverage of the two pathways, though still relatively low, 0.165.

The second pathway, $i * S$, indicates that organizations without an inclusive identity focus and with larger resources will also gain high coverage. This pathway has a consistency of 0.85 and coverage of 0.05.

The full solution, $S * (P + i)$, has an overall consistency of 0.905 and coverage of 0.218, which, again, is not very high. As with the complex solution, S is a necessary condition to high coverage.

3.5. **The Intermediate Solution.**

```
--- INTERMEDIATE SOLUTION ---
frequency cutoff: 4.000000
consistency cutoff: 0.920410
Assumptions:
s (present)
t (present)
p (present)
d (present)
i (present)
a (present)

              raw       unique
            coverage    coverage   consistency
           ----------  ----------  ----------
s*p        0.164582    0.164524    0.912729
s*~i*a     0.053190    0.053133    0.883043
solution coverage: 0.217715
solution consistency: 0.905272
```

Since the assumptions made in the parsimonious solution may not be justified, we can calculate the intermediate solution. The intermediate solution distinguishes between "easy" and "difficult" assumptions, and only includes "easy" remainders when simplifying the solution. The software asks us for these easy assumptions when it calculates the intermediate solution — when we set all causal conditions to present, we were telling the software that we believe that it is the presence of the causal conditions that lead to the outcome. So the software will use remainders that allow it to eliminate absence terms from the complex solution instead of ones that eliminate presence terms.

Our solution here is similar to the parsimonious solution, except that the second pathway requires the organization to be focused on AIDS policy. So overall, the solution, $S*(P+A*i)$, has a consistency of 0.905 and coverage of 0.218.

No solution type is necessarily better than any other. To reiterate, the complex solution makes no assumptions, the parsimonious solution uses both easy and difficult assumptions to simplify, and the intermediate only uses easy assumptions. This means that typically you'll want to present the intermediate solution, as it will be simpler than the complex solution but not involving assumptions that can't be justified. However, the simplification procedures used by the software in generating the intermediate and parsimonious solution might inadvertantly eliminate an important causal condition from the solution, in which case you may be best going with the coplex solution, or simplifying the solution manually.