Subset/Superset Analysis

The subset/superset procedure offers a way to both test a given causal recipe and to dissect it. The procedure assesses the consistency and coverage of a user-specified recipe, as well as the consistency and coverage of all subsets of ingredients in the recipe. It is called the subset/superset procedure because each subset of ingredients constitutes a superset of the cases included in the initial recipe. For example, the set of males is a superset of the set of the set of red-headed males. More generally, the fewer the ingredients in a recipe, the larger the set of cases it embraces.

The standard fsQCA procedure is truth table analysis, where the goal is to build a logical statement describing the patterns found in the data. The usual result is a specification of the different combinations of causal conditions linked to an outcome. An important part of truth table analysis is the consideration of “remainders”—the combinations of causally relevant conditions that lack empirical instances. The three different ways of treating these remainders provide the basis for the derivation of the “complex,” ”parsimonious,” and “intermediate” solutions to the truth table. Truth table analysis has an inductive quality, because these statements are constructed from the evidence summarized in the truth table.

Subset/superset analysis, by contrast, has a deductive quality. The causal "recipe" is supplied in advance by the researcher. The goal of subset/superset analysis is to assess not only the recipe as formulated by the researcher, but also to assess all possible subsets of ingredients specified in the recipe.

For example, a researcher might speculate that causal recipe a~bc~d is sufficient for outcome y (i.e., the speculation is that membership in this recipe is a consistent subset of membership in y). Notice that the researcher must specify the **directionality** of each causal condition. That is, the researcher must specify whether it is the causal condition (as calibrated in the data spreadsheet) that is linked to the outcome or its negation.

The subset/superset routine in fsQCA assesses not only the initial recipe, as specified, but also all possible subsets of the relevant conditions.

The number of recipes assessed in a single subset/superset analysis is $2^k - 1$, where k is the number of conditions in the initial recipe. For a four-condition initial recipe, the procedure assesses a total of 15 formulations:

the initial recipe (a single combination of four conditions)

a~bc~d

four three-condition recipes:

a~bc
a~b~d
ac~d
~bc~d

six two-condition recipes:

a~b
ac
a~d
~bc
~b~d
c~d

four one-condition recipes:

a
~b
c
~d

For each recipe tested, fsQCA reports the consistency and coverage of the recipe, with respect to its subset relation with the outcome. In other words, the program assesses the causal sufficiency of each recipe—do cases with this recipe constitute a fuzzy subset of cases with the outcome? If consistency is high (e.g., at least 0.80), it is reasonable to evaluate coverage. If consistency is low, then the calculation of coverage is meaningless.

Notice what is **not** assessed by the subset/superset procedure. In the truth table procedure, if the researcher specifies causal conditions a, b, c, and d as relevant, then the program assess all 16 presence/absence configurations of these four conditions:

~a~b~c~d	~a~b~cd	~a~bc~d	~a~bcd
~ab~c~d	~ab~cd	~abc~d	~abcd
a~b~c~d	a~b~cd	***a~bc~d***	a~bcd
ab~c~d	ab~cd	abc~d	abcd

The subset/superset procedure, by contrast, starts with only one of these sixteen (e.g., the one that is in bold italics) and then explores simpler versions of this recipe.

Notice also that when using the truth table procedure the movement to a simpler formulation (e.g., to the three-condition recipe a~bc) requires that the two four-condition recipes included in this recipe (i.e., a~bc~d and a~bcd) both “pass” the test for sufficiency (or its equivalent via counterfactual analysis). The testing of simpler recipes using the subset/superset procedure, by contrast, is mechanical and does not require that all the component configurations pass some sort of sufficiency test.

It is also important to understand that because the subset/superset procedure automatically tests all subsets of causal conditions, it is completely blind to the problem of limited diversity and to the importance of counterfactual analysis.

Recall that the truth table procedure's three solutions results from different treatment of the "remainder" configurations (combinations of causal conditions that lack empirical instances). The complex solution bars the use of remainders. The parsimonious solution uses any remainders that it can, regardless of whether they might be considered "easy" or "difficult" as counterfactual cases (from the perspective of theoretical and substantive knowledge). The intermediate solution uses only those remainders that are considered "easy" counterfactuals.

The subset/superset procedure uses remainders in a knowledge-blind manner. That is, like the derivation of the parsimonious truth table solution, there is no check on the use of remainders that might be considered difficult counterfactuals. Suppose, for example, that a researcher using the subset/superset procedure specifies the recipe a~bc~d. As part of the analysis, the subset/superset procedure will also automatically test a~bc. But suppose there are no cases of a~bcd (i.e., no cases with greater than 0.5 membership in this combination), and our theoretical knowledge tells us that a~bcd is a difficult counterfactual. From the perspective of the truth table procedure, the test of the sufficiency of a~bc is not warranted.

For this reason, is always best to view the subset/superset procedure as exploratory. Perhaps, based on case knowledge, a~bc might be considered a better starting recipe than a~bc~d, and the whole issue of the difficult counterfactual a~bcd, from this perspective, might be seen as a useless distraction.

As always, it is important to remember that the purpose of both procedures, the truth table procedure and the subset/superset procedure, is to facilitate the interaction between ideas and evidence in the production of social scientific representations of empirical evidence.

An Example of the Subset/Superset Procedure

Cases: European democracies between WWI and WWII
Outcome: democratic breakdown
Causal conditions: low development, low industrialization, low urbanization, political instability

Here's the initial output:

Recipe	Consistency	Coverage
~urbanfz*~develfz*~indusfz*~stablfz	0.914934	0.510548
~urbanfz*~develfz*~indusfz	0.885675	0.678270
~urbanfz*~develfz*~stablfz	0.912249	0.526371
~urbanfz*~indusfz*~stablfz	0.909927	0.522152
~develfz*~indusfz*~stablfz	0.897163	0.533755
~urbanfz*~develfz	0.890306	0.736287
~urbanfz*~indusfz	0.727072	0.694093
~urbanfz*~stablfz	0.913765	0.581224
~develfz*~indusfz	0.865110	0.703586
~develfz*~stablfz	0.894017	0.551688
~indusfz*~stablfz	0.889845	0.545359
~urbanfz	0.673859	0.856540
~develfz	0.837472	0.782700
~indusfz	0.708376	0.722574
~stablfz	0.901592	0.657173

Here it is again, sorted by coverage, deleting recipes with consistency < 0.85.

Recipe	Consistency	Coverage
~urbanfz*~develfz	0.890306	0.736287
~develfz*~indusfz	0.865110	0.703586
~urbanfz*~develfz*~indusfz	0.885675	0.678270
~stablfz	0.901592	0.657173
~urbanfz*~stablfz	0.913765	0.581224
~develfz*~stablfz	0.894017	0.551688
~indusfz*~stablfz	0.889845	0.545359
~develfz*~indusfz*~stablfz	0.897163	0.533755
~urbanfz*~develfz*~stablfz	0.912249	0.526371
~urbanfz*~indusfz*~stablfz	0.909927	0.522152
~urbanfz*~develfz*~indusfz*~stablfz	0.914934	0.510548

A key issue is parsimony: is a more parsimonious solution necessarily better if the less parsimonious solution has about the same coverage?

Sometimes more parsimonious recipes might be preferred; sometimes less parsimonious solutions might be preferred. Not only is it easier to connect *less* parsimonious recipes to empirical cases, it is important to recall that parsimony is often achieved at the expense of necessary conditions. Recall that consistency (as computed above) is an evaluation of the degree to which a causal combination constitutes a subset of the outcome. If this causal combination includes a necessary condition, then (more than likely) this condition *never* supplies the minimum value. After all, its scores are generally ≥ y. Thus, removing the necessary condition from a combination is unlikely to contribute to inconsistency.

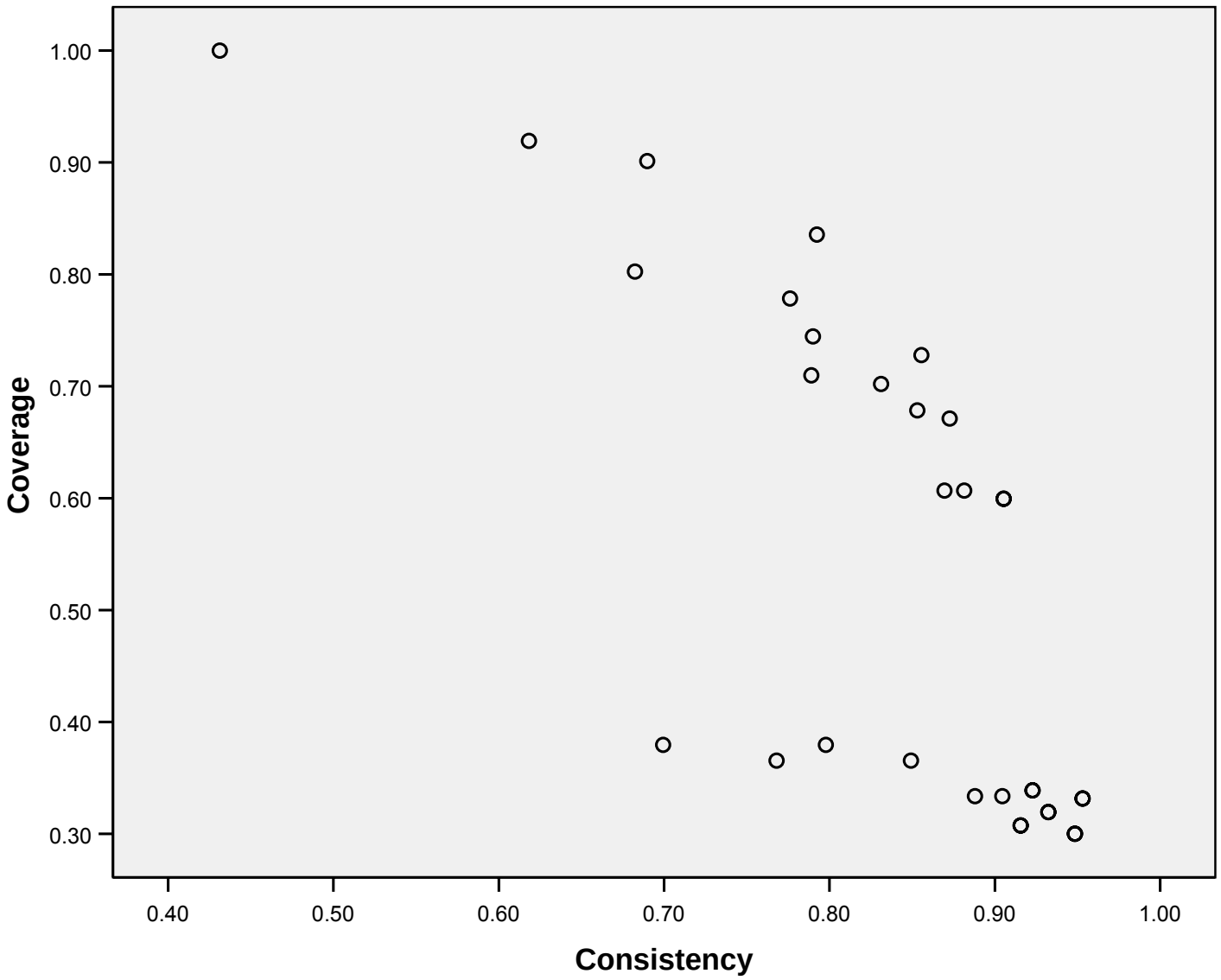
To be more precise, if the consistency of abc as a subset of y is 0.90, and c is a necessary condition, then the consistency of ab will also be about 0.90. Condition c never supplies the minimum because c is ≥ y and abc is ≤ y. Parsimony would dictate selecting ab as the best recipe; common sense dictates the selection of abc.

The next slide shows a good example of this issue. The best solution is, of course, one that maximizes both consistency and coverage. Consistency should be as close to 1.0 as possible; otherwise, the coverage calculation is meaningless. In the spreadsheet that follows, a three-condition recipe and a four conditions recipe have the same coverage and consistency. (The four condition recipe includes a necessary condition.) Which one is better? From a strict “parsimony” viewpoint, having fewer conditions is better. But from both a logical and an “understand-your-cases” viewpoint, the four-condition recipe is superior. (From a logical viewpoint, it is a bad idea to drop a condition that might be considered necessary.)

In the following spreadsheet, causal combinations with both consistency scores greater than 0.85 and coverage scores greater than 0.60 are in bold type.

Combination	Consistency	Coverage
1. press urban hard lib activ	0.95	0.30
2. press urban hard lib	0.95	0.33
3. press hard lib activ	0.95	0.30
4. press urban lib activ	0.93	0.32
5. urban hard lib activ	0.92	0.31
6. press urban hard activ	**0.91**	**0.60**
7. press hard lib	0.95	0.33
8. press lib active	0.93	0.32
9. urban hard lib	0.92	0.34
10. hard lib active	0.92	0.31
11. press hard activ	**0.91**	**0.60**
12. urban lib activ	0.90	0.33
13. urban hard active	**0.88**	**0.61**
14. press urban hard	**0.87**	**0.67**
15. press urban active	**0.86**	**0.73**
16. press urban lib	0.85	0.37
17. hard lib	0.92	0.34
18. lib active	0.89	0.33
19. hard active	**0.87**	**0.61**
20. urban hard	**0.85**	**0.68**
21. press hard	0.83	0.70
22. urban lib	0.80	0.38
23. press urban	0.79	0.84
24. press active	0.79	0.74
25. press lib	0.77	0.37
26. urban active	0.78	0.78
27. hard	0.79	0.71
28. lib	0.70	0.38
29. urban	0.69	0.90
30. active	0.68	0.80
31. press	0.62	0.92

Here's a plot of the consistency and coverage scores from the spreadsheet. It's clear that there are two groups of coverage scores, those that include political liberalization (lib) as a causal condition (the low-coverage group) and those that exclude it (the high-coverage group).


Coverage
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Consistency