Article

Sociological Methods & Research
2017, Vol. 46(2) 242-251

DOI: 10.1177/0049124116672701
journals.sagepub.com/home/smr
SAGE

Aggregation Bias and the Analysis of Necessary and Sufficient Conditions in fsQCA

Bear F. Braumoeller[1]

Abstract

Fuzzy-set qualitative comparative analysis (fsQCA) has become one of the most prominent methods in the social sciences for capturing causal complexity, especially for scholars with small- and medium-*N* data sets. This research note explores two key assumptions in fsQCA's methodology for testing for necessary and sufficient conditions—the cumulation assumption and the triangular data assumption—and argues that, in combination, they produce a form of aggregation bias that has not been recognized in the fsQCA literature. It also offers a straightforward test to help researchers answer the question of whether their findings are plausibly the result of aggregation bias.

Keywords

interactions, QCA, Charles Ragin, necessary and sufficient conditions, fsQCA, aggregation bias

fsQCA has become one of the most prominent methods in the social sciences for capturing causal complexity, especially for scholars with small- and

[1] Department of Political Science, The Ohio State University, Columbus, OH, USA

Corresponding Author:
Bear F. Braumoeller, Department of Political Science, The Ohio State University, 2168 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, USA.
Email: braumoeller.1@osu.edu

medium-N data sets. The volume that introduces fsQCA, Charles Ragin's *Fuzzy-Set Social Science*, has been cited 1,895 times since its publication in 2000. The reasons are straightforward: Given its foundation in Boolean algebra, it captures the logic of complexity quite well, it is versatile enough to handle a wide range of problems, it has been implemented in easy-to-use software, and the assumptions required to arrive at conclusions are mostly sensible (though see Hug 2013; Krogslund, Choi, and Poertner 2015; Braumoeller 2015, for dissenting views of varying strength).

While the technique's assumptions may be individually sensible, however, they can be problematic in combination. This article explores two of these assumptions in fsQCA—the *cumulation assumption* and the *triangular data assumption*—and argues that, in combination, they produce a form of aggregation bias that has not been recognized in the fsQCA literature. It also offers a straightforward test to help researchers answer the question of whether their findings are plausibly the result of aggregation bias.

Necessary and Sufficient Conditions

fsQCA analyses of the complex causes of necessary and sufficient conditions (Ragin 2000:234-38) incorporate two definitional assumptions that are individually innocuous but jointly worrisome. The first is what I will call the cumulation assumption—the assumption that the value of the intersection of two fuzzy sets is equal to the minimum of separate memberships in the fuzzy sets that define them, and the value of the union of two sets is equal to the maximum of those sets' separate fuzzy membership scores. For example, if a country's membership in the set "democratic states" is 0.7 and its membership in the set "industrialized states" is 0.5, its membership in the set "industrialized democracies" will be equal to 0.5—it cannot, logically, be more of an industrialized democracy than it is an industrialized state. Its level of industrialization is the weak link that defines its membership in the joint set. By the same token, that state's membership in the set of democratic *or* industrialized states would be the maximum of the two individual scores or 0.7.

The second assumption, which I will refer to as the "triangular data assumption," posits that necessary- and sufficient-condition relationships between conditions and outcomes are evidenced by a triangular data pattern in an ordinary scatterplot (Figure 1). If a unit of X is necessary for a unit of Y, no (or, given the possibility of measurement error, very few) observations should be observed above the line $X = Y$. Similarly, if a unit of X is sufficient for a unit of Y, no (or very few) observations should be observed below

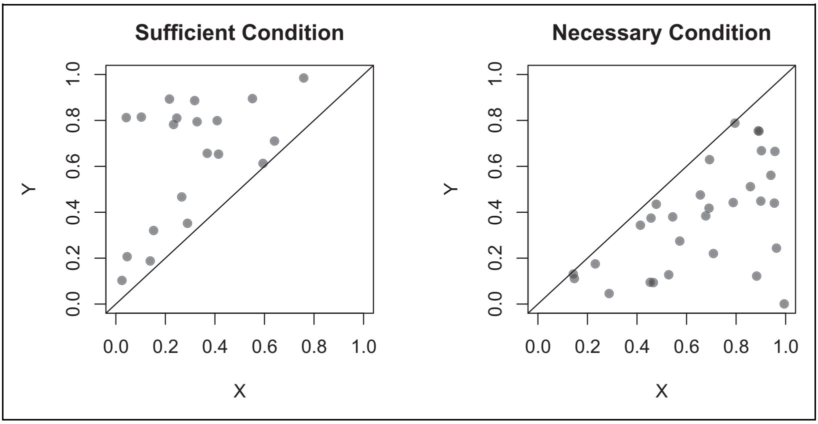

Figure 1. Triangular patterns of data, indicative of necessary and sufficient conditions.

the line $X = Y$. The result, in both cases, is a set of observations shaped roughly like a triangle.

Neither of these assumptions is problematic in and of itself.[1] Indeed, the unit range of fuzzy-set membership makes it an ideal candidate for bivariate necessary- and sufficient-condition analysis. The problem arises in the combination of the two: The minimization and maximization algorithms used to calculate the intersections and unions of sets create aggregation bias, which increases the risk of false positives.

Aggregation Bias

To illustrate this problem, I use Monte Carlo simulations to show that the minimization algorithm, used on randomly generated data, produces patterns that resemble a sufficient condition. (The same is true for the maximization algorithm and necessary conditions, as intuition might suggest.) To begin, I generate six simulated conditions $X_1 - X_6$ and a simulated outcome Y, each comprising 50 random numbers drawn from a uniform distribution. Each of the simulated conditions represents the membership of 50 notional observations in a hypothetical fuzzy set, and the minimum of any number of these conditions' values in a given row represents that observation's degree of membership in the subset of those fuzzy sets. Because the conditions are uncorrelated by construction, any test of any combination of $X_1 - X_6$ as a sufficient condition for Y should fail or at least not produce a false positive in more than 5 percent of all cases.

In fact, defining membership in a fuzzy subset as the minimum of memberships in fuzzy sets systematically drives subset membership downward, on average. This is as it should be: As more fuzzy sets are added on to the list of those used to create the subset, there are more and more opportunities for a new weak membership to define the minimum and therefore the joint fuzzy-subset membership.

Nevertheless, that tendency has an unforeseen and undesirable effect on tests of sufficiency, as we can see in Figure 2. In general, aggregation tends to decrease the range of the distribution of joint fuzzy-subset memberships. Unfortunately, as that shrinking occurs, a larger and larger percentage of the observations in the data set ends up above the diagonal line at $X = Y$, and the result bears a closer and closer resemblance to a sufficient-condition relationship—despite the fact that, by construction, the conditions are totally uncorrelated with the outcome.

A Remedy: The $\hat{p}_I$ Test

The ideal remedy to this problem is not entirely obvious, but some aspects of it are straightforward. The first step to understanding what a test should look like is understanding what the data should look like under the null hypothesis—that is, if, as in these simulations, they are entirely uncorrelated. The answer here is straightforward: For sufficient conditions, on average, the value of Y should be less than the value of X_1 about 50 percent of the time; less than the values of both X_1 and X_2 about one third of the time; less than the values of X_1, X_2, and X_3 about one fourth of the time; and so on. Accordingly, the fraction of the observations above the line $X = Y$ under the null hypothesis should equal $1 - \frac{1}{n+1}$, where n represents the number of fuzzy-set memberships that are used to construct the fuzzy subset. So

$$\mathrm{H}_0 : Pr(Y > X) = 1 - \frac{1}{n+1}.$$

The calculations for necessary conditions mirror these precisely. For necessary conditions, on average, the value of Y should be greater than the value of X_1 about 50 percent of the time, greater than the values of both X_1 and X_2 about one third of the time, and so forth, leading to a parallel hypothesis for necessity:

$$\mathrm{H}_0 : Pr(Y < X) = 1 - \frac{1}{n+1}.$$

Differentiating between the data that one would expect to observe under the null hypothesis and the data that one would expect to see under a

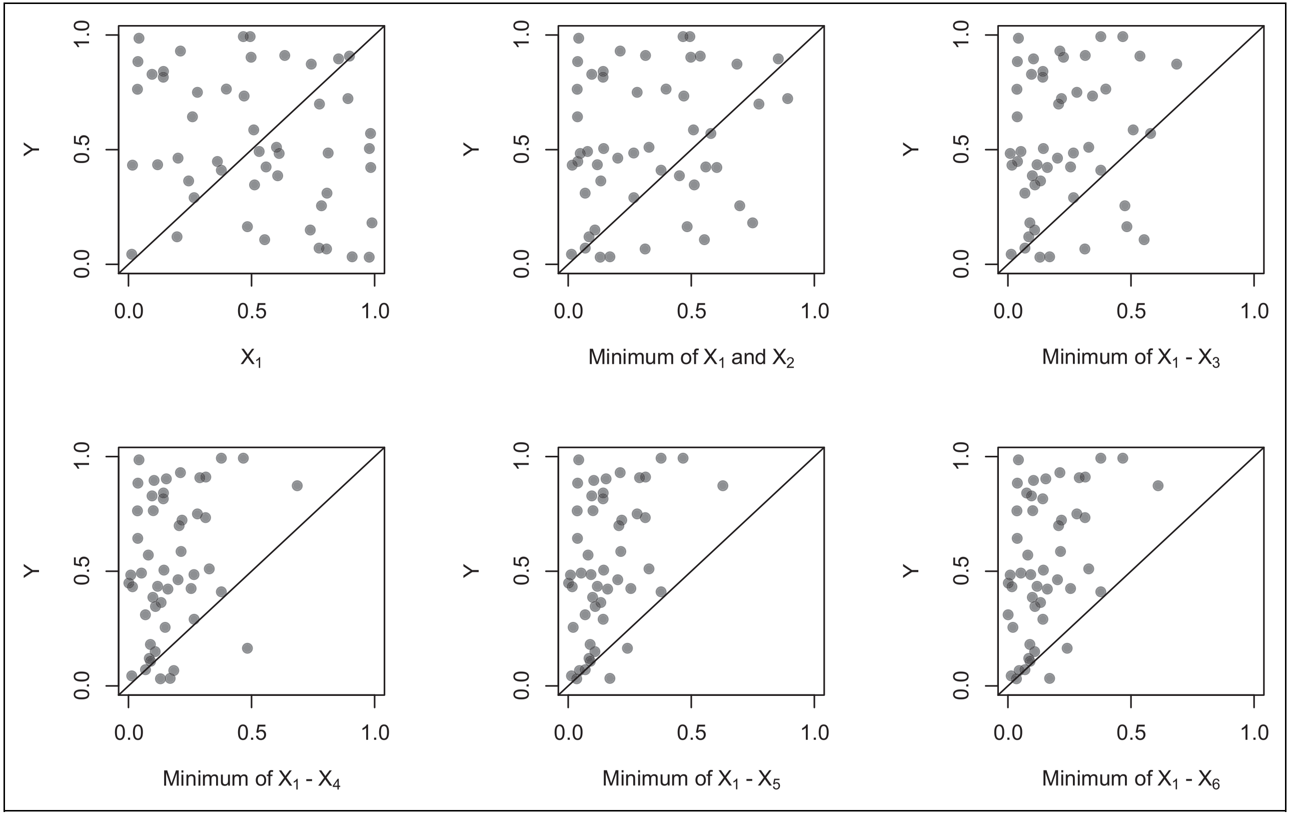

Figure 2. The impact of minimization as a measure of the intersection of fuzzy subsets on tests of sufficiency in random data.

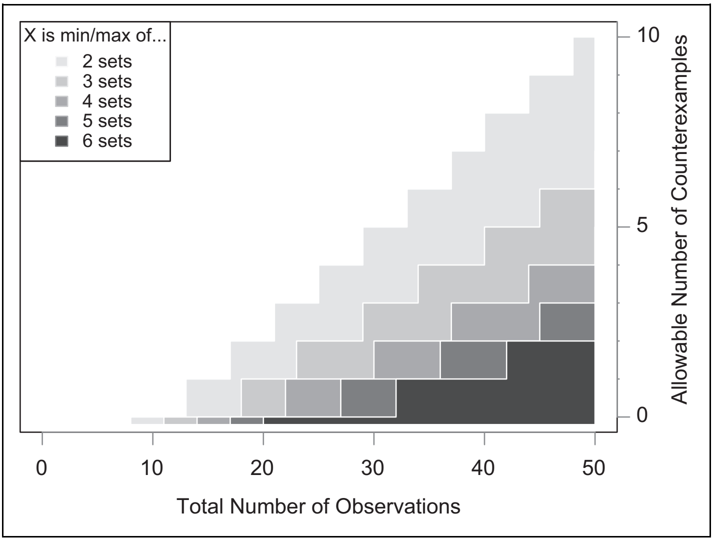

Figure 3. Maximum number of acceptable counterexamples by number of observations and number of set intersections.

hypothesis of necessity or sufficiency is a bit trickier. Ideally, one would utilize a test that incorporates the two data generating processes, but the data generating process for a sufficient condition is not as well specified as we might like. Literally, any distribution of points above the line $X = Y$ would satisfy the fsQCA definition of a sufficient condition, regardless of how concentrated, dispersed, or skewed they might be, and the same is true for points below the line $X = Y$ and necessary conditions.

The silver lining here is the fact that, since the precise distribution of observations contains no useful information, we can ignore it and focus entirely on the proportion of outliers. Given that information, the $\hat{p}_I$ test of necessary conditions proposed by Braumoeller and Goertz (2000) can be of considerable use. The gist of the test, applied to this context, is that, in order to reject the null hypothesis of no association when X is a fuzzy-subset membership derived from the intersection or union of n fuzzy-set memberships, a 95 percent one-sided binomial confidence interval around the observed proportion of confirming, or consistent, cases should exclude $p = 1 - \frac{1}{n+1}$.

Figure 3 illustrates the relationship that the test implies between the total number of observations and the allowable number of counterexamples for the union or intersection of anywhere from two to six sets. In the case of the lower right panel of Figure 2, which contains 50 observations and a fuzzy-

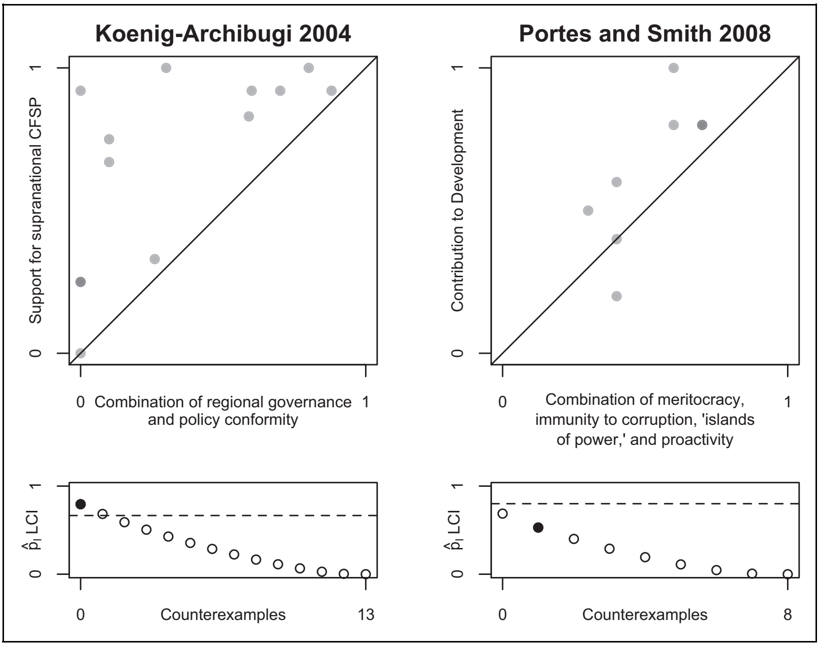

Figure 4. Replications of two studies, showing results of $\hat{p}_l$ tests.

subset membership derived from six set memberships, the three counterexamples would be insufficient to reject the null hypothesis: The one-sided 95 percent confidence interval at 0.852 would fail to exclude the null hypothesis of $p = 1 - \frac{1}{6+1} = 0.86$.[2]

Examples

Two examples of medium-*N* fsQCA studies with superficially strong results will help to demonstrate how the test works. In the first study, Koenig-Archibugi (2004) examines European states' attitudes regarding a common foreign and security policy (CFSP). The author finds that the combination of preference for regional governance at the domestic level and "policy conformity" (policy preferences not far from the norm) is jointly sufficient to ensure support for a CFSP. In the 13 cases examined, there is not a single exception to this generalization.

The left-hand column of Figure 4 illustrates this relationship. The top graph plots the data relative to the line $Y = X$. The bottom chart shows the lower bound of the one-sided 95 percent binomial confidence interval around the observed proportion of confirming cases, as the hypothetical number of counterexamples ranges from 0 to 13. The dashed horizontal line at $Y = 1 - \frac{1}{2+1} = 0.67$ represents the value that the lower confidence interval must exclude if we are to be confident that the results differ significantly from those that we would expect due to aggregation bias. The black dot at 0 counterexamples indicates that the confidence interval does, in fact, exclude 0.67. The remaining dots illustrate the fact that, given the n of 13 and the fact that only two sets are aggregated, the study would also have passed the test with a single counterexample—but not with two or more.

By contrast, a study of the determinants of the development of effective institutions by Portes and Smith (2008), illustrated in the right-hand column of Figure 4, reports results that are very plausibly the result of aggregation bias. With only eight observations and a single counterexample, it would be difficult to rule out aggregation bias even under the best of circumstances. Given that the causal condition reflects the intersection of four sets—meritocracy, immunity to corruption, the absence of "islands of power" within the government, and proactivity or the ability of the government to engage other actors—even perfect results with eight observations could not exclude $Y = 1 - \frac{1}{4+1} = 0.80$.

As these results, and Figure 3, suggest, the $\hat{p}_I$ test is tough but fair. It is difficult to distinguish the results of studies with a very small n and/or those in which the causal condition comprises the aggregation of many sets from the results of aggregation bias—which is just as it should be. At the same time, the test allows a reasonable number of counterexamples in studies with a slightly larger number of cases or a more modest number of sets.

Conclusion

This brief discussion has outlined a heretofore-unappreciated danger in fsQCA—the problem of aggregation bias—and has offered an extremely simple test to evaluate the argument that aggregation bias is responsible for a given set of findings.

To be very clear, this test is *not* a test of necessity or sufficiency. Rather, it is a test to determine whether the distribution of the observed data is sufficiently different from the distribution implied by the null hypothesis to warrant rejection of the latter. Put more succinctly, it is a test of the hypothesis that aggregation bias, rather than a meaningful relationship among variables, produced the data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. It might well be worth relaxing the assumption that a unit of X is necessary/sufficient for a unit of Y, and instead exploring the question of how much X is necessary/sufficient for how much Y, but it would be too much of a digression to do so here.
2. A brief aside: One anonymous reviewer suggests that it should be used "after each minimization in order to test whether the various paths (i.e., combinations of X that are sufficient for Y) highlighted show enough variations to be interpreted meaningfully." My own inclination is to use this test on the final, reduced-form Boolean formula only, but I can see the value of iterated testing as a heuristic device.

References

Braumoeller, Bear F. 2015. "Guarding Against False Positives in Qualitative Comparative Analysis." *Political Analysis* 23:471-87.

Braumoeller, Bear and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44:844-58.

Hug, Simon. 2013. "Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference." *Political Analysis* 21:252-65.

Koenig-Archibugi, Mathias. 2004. "Explaining Government Preferences for Institutional Change in EU Foreign and Security Policy." *International Organization* 54: 137-74.

Krogslund, Chris, Donghyun Danny Choi, and Mathias Poertner. 2015. "Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA." *Political Analysis* 23:21-41.

Portes, Alejandro and Lori D. Smith. 2008. "Institutions and Development in Latin America: A Comparative Analysis." *Studies in Comparative International Development* 43:101-28.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago, IL: University of Chicago Press.

Author Biography

Bear F. Braumoeller is an associate professor in the Department of Political Science at the Ohio State University. He is grateful to Austin Knuppe and Andrew Rosenberg for their research assistance.