Battling Methodologies: Rare Events and Methodological Choices*

Christopher K. Butler
University of New Mexico

Kendra L. Koivu
University of New Mexico

October 21, 2014

PRELIMINARY DRAFT.
DO NOT CITE WITHOUT PERMISSION.

Abstract

Methodological diversity is praised in some circles for providing a larger set of tools to the researcher, but it is possible that different tools provide different answers to the same question. How do our methodological choices affect the answers we get? One area of methodological diversity that provides some similarity to large-N statistical analysis is boolean analysis. We analyze data from both approaches. We start by examining simulated rare-events data sets following logistic or boolean logics, analyzing them separately, and then comparing results. When the data generation process is known, we find that both statistical and boolean approaches provide largely the same results. When the data generation process is not known, however, the results are dependent on the choices researchers make. We then examine civil-war onset using both approaches. We find that large-N, rare-events data sets present challenges to both statistical and boolean approaches. While these challenges have been largely resolved in statistical studies, boolean approaches have yet to establish best practices. We offer suggestions from our study regarding how boolean analysis can tackle large-N, rare-events data sets. We also summarize the findings of boolean analysis regarding civil-war onset and compare these findings with extant statistical findings in the literature.

*Prepared for the annual Fall Research Conference of the Association for Public Policy Analysis & Management, November 6-8, 2014, in Albuquerque, NM. We thank Colin Hannigan for research assistance on this paper.

We examine two methodologies that can be used to study rare events: logistic regression and Boolean analysis. The impetus for this comparison is straightforward: do the two techniques give the same results when applied to the same question? The answer to this question is important to users of each technique. If the answer is "No", we would all do well to understand where the deviation arises. If the answer is "Yes, practitioners of each technique could collaborate more freely and, when the occasion arose, use the other technique to overcome an obstacle within their more accustomed technique.

In this paper, we construct simulated datasets to be analyzed by each method. The simulated data allow us to know what the real relationship is. It also provided us with an opportunity to embed problems that each technique is supposed to be good at. For example, we included U-shaped and inverted-U shaped relationships that are discoverable using logistic analysis. We also included datasets with multiple interaction terms that are discoverable using Boolean analysis. Finally, we started with an eye toward examining civil war onset using each technique. While we did not get to this component of the research agenda, we did simulate datasets having rare events (less than 10% positive cases). This allows us to examine how the techniques handle such data.

The Methodologies

In this section, we separately review what each type of methodology entails.

Logistic Analysis

Logistic regression (or logit) is one of a large set of limited-dependent variable statistical techniques. Logit is an appropriate technique if we can only observe the presence or absence of an event (such as civil war). Logistic analysis tests whether the included independent variables are correlated with the dependent variable. If they are (via significance), then we interpret the signs of their coefficients in terms of increasing (for positive coefficients) or decreasing (for negative coefficients) the likelihood of the dependent variable occurring. We can also calculate the overall likelihood (or probability) of the dependent variable occurring for a specific case given the values of the independent variables for that case. This case may be within the sample, in which case we are assessing the extent to which the model would have predicted that case correctly. If we believe that the model is generalizable, then we can use the values of the independent variables for cases not included in the model to generate predicted probabilities for those "out-of-sample" cases. This can be done by not including all cases in the analysis that generates the coefficients to assess the generalizability of the model. This can also be done using new cases to make real-time predictions, though this is seldom done in political science.

Having described the basic ideas of logistic analysis, we now describe it mathematically. The observed values of the dependent variable are 1 for the presence of the event and 0 for the absence of the event: $y \in \{0, 1\}$. The independent variables, $x_1, x_2, \cdots, x_m$, can be continuous or dichotomous variables. The estimated coefficients for each independent variable is given by $\beta_1, \beta_2, \cdots, \beta_m$ and β_0 is the estimated "constant" of the model (representing the baseline predicted probability when all independent variables equal zero). The estimated predicted probability of an event occurring is given by equation 1.

$$y^* = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}} \tag{1}$$

The predicted probabilities are also used to classify cases as as predicted to occur or not. Specifically, the discrete prediction, $\hat{y}$, is given by:

$$\hat{y} = \begin{cases} 1 & \text{if } y^* \geq 0.50 \\ 0 & \text{if } y^* < 0.50 \end{cases}$$

An observed case is considered to be correctly predicted by the model when $y = \hat{y}$.

Qualitative Comparative Analysis

Social scientists, in the pursuit of rigorous methods that enable us to have confidence in the conclusions we draw, have our hands tied. We cannot hold variables constant in a laboratory for most of the questions for which we seek answers. We usually cannot conduct randomized, or controlled, experiments. If we are fortunate, nature gives us a natural experiment, where the experimental conditions are determined by exogenous factors outside of the researchers control.[1] In this instance, two populations exist in near similar conditions, and one of the populations naturally receives a treatment that the researcher happens to be interested in knowing the effects of. However, these are observational studies that cannot be created at will.

To get around the problem of having no social science laboratory, researchers frequently turn to quantitative analysis. Statistical techniques allow social science researchers to evaluate large amounts of data to identify trends and patterns. Quantitative analysis can help us understand not only the magnitude of the causal relationship we are evaluating; it also gives us confidence that the effect in the sample population we are observing is not by chance, but rather can be applied to the universe of cases in our population. However, there are drawbacks to quantitative analysis.[2] The one that concerns us here is the problem of rare events. Quantitative analysis works best with large numbers of observations (greater than 50), but what if the events we are interested in happen only rarely? Again, the hands of the social scientist are seemingly bound.

The response within the quantitative methods camp has been to develop innovative uses of statistical analysis that can adequately handle a small- or medium-N research design. Case study researchers, on the other hand, have advocated for smaller studies that emphasize within-case analysis and paired comparisons. Then, they argue, it is only a matter of getting the case selection right. But single-case studies, rich in detail as they may be, are difficult to generalize from, and paired comparisons may still suffer from poor case selection, particularly if the universe does not provide appropriately comparable cases. There is, however, another approach. Qualitative comparative analysis (QCA) is uniquely suited for medium-N research designs. They have the added advantage of disentangling causal heterogeneity, limited diversity, and equifinality, all three of which are reasonable expectations given the complexity of the social world.

QCA, also known as set-theoretic methods or configurational comparative methods (CCM), is based on Boolean algebra and formal logic and allows for the systematic comparison of cases. QCA and set-theoretic methods can be defined as "approaches to analyzing social reality in which (a) the data consists of set membership scores; (b) relations between social phenomena are modeled in terms of set relations; and (c) the results point to sufficient and necessary conditions and emphasize causal complexity in terms of INUS and SUIN causes."[3] Causal complexity entails equifinality, conjunctural causation and causal asymmetry. Equifinality means that there are potentially many paths to one outcome. Conjunctural causation refers to the way in which factors exert a causal effect on an outcome, i.e. in combination with other specified factors. This is sometimes referred to as *multiple conjunctural causation*, "a nonlinear, non-additive, non-probabilistic conception that rejects any form of permanent causality and that stresses equifinality (different paths can lead to the same outcome), complex combinations of conditions, and diversity."[4] Causal asymmetry implies that the presence and absence of an outcome may have different explanations. Whereas a symmetric understanding of $X \rightarrow Y$ should lead us to conclude that $\sim X \rightarrow \ \sim Y$, an asymmetric understanding of $X \rightarrow Y$ would not permit that conclusion. The outcomes Y and $\sim Y$ could have entirely different causal conditions.

Theoretically, there are a number of reasons to expect that a set-theoretic approach, with its emphasis on causal complexity, is appropriate for civil war onset. It is reasonable to assume that cases of civil war onset can be characterized by *equifinality* (multiple causal paths), not *unifinality* (one causal path). Simply put, civil wars may occur in different places for different reasons, and QCA/set-theoretic approaches are uniquely suited for detecting this. It is also reasonable to assume that the causal conditions of civil wars are best characterized as *conjunctural*, not *additive*. Causal additivity is when a cause can be said to have

[1] The study of the 1854 cholera outbreak in London is considered the first natural experiment.

[2] Cite those chapters from RSI (Brady, Bartels and Rogowski).

[3] Schneider and Wagemann 2012, p. 6.

[4] Rihoux and Ragin, p. 17.

an independent effect on the outcome: "the strength of its effect is not altered by the presence or absence or level of other causes."[5] Statistical analysis is a powerful tool for identifying a set of causal variables that have an independent effect on the outcome and explain as much of the variation as possible. However, if a variables causal effect is not independent but dependent on the presence or absence of other variables (i.e., conjunctural), then a set-theoretic approach is more appropriate.[6] It may be that some proposed causal conditions of civil war onset, such as ethnolinguistic fragmentation, negative change in GDP, and an autocratic state do not by themselves lead to civil war, but only when in the presence of each other. Additionally, civil war onset may best be characterized by *causal asymmetry*, not *causal symmetry*. The causal conditions that explain the outcome of civil war onset may be entirely different from the set of conditions that explain non-civil war onset. For these reasons, we have good reasons to expect, from a theoretical standpoint, that a set-theoretic approach might be appropriate for studying civil war onset.

However, the problem *rare events* poses for statistical analysis is potentially confounding for QCA as well. The reason for this is the nature of the subset relationship between the causal factor(s) and outcome of rare events that QCA is meant to evaluate. A rare event is one in which a particular phenomenon could potentially occur in a large number of cases, but happens significantly less frequently. Civil war onset is considered a rare event because, given the large number of cases in which they could occur, civil wars occur only rarely. For evaluating the potential subset relationship between the putative causes of civil war onset and the outcome of civil war, this poses a problem because the subset of civil wars will be extremely small, relative to the superset of causal conditions. This can effect coverage and consistency scores, the means by which we interpret QCA results.

Sufficient Conditions

The information QCA gives us concerning the subset relationship take the form of *coverage* and *consistency* scores. Consistency scores for sufficient conditions indicate the degree to which the posited subset relationship exists. Consistency scores for sufficiency must be adequate before moving to interpret coverage scores. Consistency scores are derived by summing all cases where both the outcome and cause are present and dividing them by the sum of all cases where the outcome is present. If there are many cases that display the cause but not the outcome, but relatively few cases that display both the cause and the outcome, the consistency will be low. This represents a subset relationship where the causal condition set is not fully subsumed by the outcome set, and is expressed as:

$$\text{Consistency of } X \text{ as a sufficient condition for } Y = \frac{\text{Number of cases where } X = 1 \text{ and } Y = 1}{\text{Number of cases where } X = 1} \tag{2}$$

We have reason to expect that consistency scores for rare events will be naturally low because the numerator in the equation above should be much larger than the denominator. This makes consistency scores difficult to interpret. While Schneider and Wagemann indicate that the preferable threshold for sufficient conditions should 0.75, this is not practical for rare events because the outcome occurs only rarely, relative to all the causal conditions.[7] To address this, two techniques will be used. First, sufficiency tests will be run on the data sets with all cases using a low threshold for consistency (0.1 and 0.2). Then, sufficiency tests will be run on the data sets with only positive cases to increase the consistency scores. Results from all sufficiency tests will be compared.

Coverage scores for sufficiency indicate the relationship in size between the causal condition(s) and the outcome. Typically, when the cause and outcome sets are close in size, the relationship is considered an empirically important one. For instance, if we are evaluating 30 cases that feature outcome Y, and the causal condition X explains 27 of those cases (90%), this is considered to be a high coverage score. The coverage

[5] Ragin 2000 p. 95.

[6] While statistical analysis provides a way to introduce interaction models, these tend to be highly collinear if specific conditions are not met (Ragin 1987). Additionally, while third- and fourth-order interaction terms are rare, set-theoretic approaches often detect combinations of three or more causal conditions (Schneider and Wagemann 2012).

[7] Schneider and Wagemann 2012, pp. 123-129.

score is a measurement of the ratio between cases where the outcome is present, and cases where both the outcome and the cause are present, and is expressed as:

$$\text{Coverage of } X \text{ as a sufficient condition for } Y = \frac{\text{Number of cases where } X = 1 \text{ and } Y = 1}{\text{Number of cases where } Y = 1} \tag{3}$$

Coverage scores take into account how many cases have high membership in the outcome set (Y), but low membership in the causal condition set (X). A case with a high score on the outcome set, but low score on the causal condition set would lower the coverage score because it indicates an instance where the outcome occurred by other means. However, because equifinality is assumed in QCA, low coverage scores are acceptable, though they explain only a small portion of the outcome. However, for the purposes of interpreting coverage for generated data sets, only causal conditions with higher than 0.60 coverage scores will be included in the solution formula.

Necessary Conditions

Consistency scores for necessary conditions are a measure of the extent to which the posited subset relationship exists, and are substantively similar to consistency scores for sufficiency. In both cases, consistency scores must be evaluated before coverage scores. The degree to which the subset relationship of necessity exists is a matter of the ratio of cases where both the cause and outcome is present, and cases where the outcome is present and the cause is absent. This is expressed as:

$$\text{Consistency of } X \text{ as a necessary condition for } Y = \frac{\text{Number of cases where } X = 1 \text{ and } Y = 1}{\text{Number of cases where } Y = 1} \tag{4}$$

If the consistency score is satisfactory, the researcher should then move to determine coverage scores. For the purposes of this paper, the consistency threshold will be low, 0.60. Coverage scores for necessary conditions indicate the extent to which a necessary cause is relevant, or a non-trivial one. A trivial necessary cause is one that is present irrespective of the outcome. Air, for instance, is a trivial necessary cause of social revolutions. Coverage scores for necessity are the ratio between cases where the cause is present and cases where both the cause and outcome are present, and is expressed as:

$$\text{Coverage of } X \text{ as a necessary condition for } Y = \frac{\text{Number of cases where } X = 1 \text{ and } Y = 1}{\text{Number of cases where } X = 1} \tag{5}$$

Coverage scores for necessity are similar to consistency scores for sufficiency. Rare events, however, will naturally have low coverage scores because a rare class of events is already a small subset of the cases evaluated. This makes the coverage score difficult to interpret. For the purposes of this paper, coverage scores will be reported, but they are expected to be so small as to be uninterpretable.

Simulated Data and Analysis

To compare logistic and QCA analyses in a controlled setting, we generated random datasets with specified characteristics. Each dataset has 1000 observations, five independent variables (uniformily distributed on $[0, 1]$), and one dichotomous dependent variable. The choice to have the independent variables bounded between zero and one was made to facilitate the use of QCA without then having to make more arbitrary decisions regarding transforming the independent variables. While logistic analysis can have continuous variables with any range, QCA requires its causal factors to be on the unit interval.

Each dataset also represents a rare-events dataset in which the dependent variable has a low occurrence, with an average mean of 5.6% across all datasets. (The lowest mean occurrence of the dependent variable is 4.9% and the highest mean is 6.7%.)

The formula by which the dependent variable was generated (i.e., the link function) represents the distinguishing feature of each dataset. That is, an intermediary linear equation (of the general form $t =$

$\alpha + \beta x + \varepsilon$) was generated that represented the "true" relationship among the independent variables and the dependent variable. This linear equation was then transformed to a probability following the logistic function in equation 6.

$$y^* = \frac{1}{1 + e^{-t}} \tag{6}$$

The cases were then classified as present or absent for y by comparing whether y^* was above or below one half:

$$y = \begin{cases} 1 & \text{if } y^* \geq 0.50 \\ 0 & \text{if } y^* < 0.50 \end{cases}$$

For several of the link functions, other temporary variables were created to represent interaction terms or induce U-shaped or inverted-U relationships. The error term of each of the link function was normally distributed with mean zero: $\varepsilon \sim N(0, 1.50)$. In all cases, the only variables kept in each final dataset were an id variable, x_1 through x_5, and y.

Equations 7 through 14 are the link functions associated with datasets 1 through 8. Link functions t_1, t_2, and t_6 represent standard monotonic, additive relationships in which there are no interaction terms. Link functions t_3 and t_8 represent curvilinear relationships in which the probability of occurrence is highest for middle values of x_2 (as in t_8) or for extreme values of x_2 (as in t_3).[8] Link functions t_4, t_5, and t_7 represent relationships with simple multiplicative interaction terms.

$$t_1 = -4.00 - 2.00x_3 + 4.00x_5 + \varepsilon \tag{7}$$

$$t_2 = -4.00 - 2.00x_2 + 4.00x_4 + \varepsilon \tag{8}$$

$$t_3 = -4.00 + 4.00\left(x_2^2 + (1 - x_2)^2 - 0.25\right) + \varepsilon \tag{9}$$

$$t_4 = -4.50 + 2.00x_4 + 4.00\left(x_3 x_5\right) + \varepsilon \tag{10}$$

$$t_5 = -3.50 + 4.00\left(x_2 x_3\right) - 3.00\left(x_4 x_5\right) + \varepsilon \tag{11}$$

$$t_6 = -4.00 + 3.00x_2 - 5.00x_3 + 6.00x_4 - 6.00x_5 + \varepsilon \tag{12}$$

$$t_7 = -4.00 + 2.00x_2 + 3.00x_5 - 4.00\left(x_2 x_5\right) + \varepsilon \tag{13}$$

$$t_8 = -4.75 - 3.00x_5 + 5.50\left(1.25 - x_2^2 - (1 - x_2)^2\right) + \varepsilon \tag{14}$$

One author created the link functions and generated the datasets. The other author then applied QCA to each dataset without knowing what the underlying relationships were. A graduate assistant applied the likelihood-ratio model-reduction method to the datasets, also without knowing what the underlying relationships were.[9]

In constructing the link functions, we deemed it important to include the following types of relationships. Some of the models include simple positive and negative relationships between the independent variables and the dependent variable corresponding to the

[8]The particular functional form for these curvilinear relationships was chosen such that a U-shaped (or inverted-U shaped) function would exist within the range $[0, 1]$ and domain $[0, 1]$ as well as have values that cross the $y = 0.50$ line twice (once for an x value below 0.50 again for an x value above 0.50). The resulting functions represent only one kind of U- or inverted-U shaped function; "J" functions are not represented in this analysis.

[9]The authors thank Colin Hannigan for his research assistance on this project.

Analysis Methods

Logistic Analysis

To an outside researcher, applying logistic regression for our battling methodologies presents a formidable problem. Essentially, the researcher is presented with the eight randomly generated data sets that have the same number of variables and of observations and then asked to identify the "best" model for each data set. While the dependent variable is clearly identified, all the other variables are equally interchangeable, having the same descriptive statistics (in that their means, etc., are statistically indistinguishable). (See Table 1 for the summary statistics of the data sets.) Additionally, none of the variables have theoretical meaning.

[Table 1 about here]

The closest technique that we have to grapple with this problem is that of model reduction using likelihood-ratio tests.In this technique, all independent variables are identified and used to construct an "unconstrained" or full model. For our purposes, this full model includes testing for interaction terms and curvilinear relationships.[10] The full model that was the starting point for the logistic analyses is given in equation 15.

$$y = f\left(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_2^2 + \beta_7 x_5 x_3 + \beta_8 x_5 x_4 + \beta_9 x_2 x_3 + \beta_{10} x_2 x_5\right) \tag{15}$$

After running this unconstrained model, the significance of the independent variables was examined to determine which would be kept for the constrained model. Our decision rule regarding this was to keep all variables with a p-value < 0.10. An additional caveat was that all component variables of an interaction term would be included regardless of their significance.[11] A likelihood-ratio (LR) test was then conducted comparing the models. If the LR test was significant, then the full model was identified as the best model. If the LR test was insignificant, then the full model was discarded as the best model. In this case, the unconstrained model was examined for further reduction following the same procedures as above.[12]

Qualitative Comparative Analysis

On each of the eight generated data sets, all variables were first converted to crisp sets.[13] Then, in an effort to reconstitute the population and prevent low and uninterpretable consistency and coverage scores, an additional data set with only positive cases was created for each original data set. This gave us our eight original, all-cases data sets and eight additional, positive-on-outcome data sets to test. Testing all cases led to predictably low consistency scores for sufficiency and coverage scores for necessity. This was supplemented with a positive-on-outcome approach to improve scores on consistency and coverage. The all-cases approach and positive-on-outcome approach required different decisions concerning frequency and consistency cut-off points when editing the truth tables. For the all-cases approach, consistency scores were low and there were cases in each property space. Given that there were (quite a few) cases in every property space, editing the truth table based on a frequency cut-off proved unsystematic. Instead, consistency scores were used as

[10] QCA examines all possible interactions as part of its technique. Such a model in a single logit would be very large and fraught with colinearity issues. Thus, we only included interaction terms that were present in at least one of the simulated data sets. Similarly, QCA can test for curvilinear relationships only if there is a suspicion (theoretically or empirically) that such a relationship exists; the QCA analyses were informed that there may be a curvilinear relationship involving x_2 in some of the models, but which models were not revealed.

[11] For example, if x_2x_3 had $p < 0.10$, then x_2 and x_3 would be included regardless of their p-values.

[12] This only resulted in further reduction in one case, involving the data set associated with link function t_4.

[13] For all generated data sets, the outcome variable was already crisp. When this is the case, fuzzy causal variables could cause the researcher to overestimate causal effects because sufficiency and necessity relies on determining whether scores on the causal variables are less than or equal to (in the case of sufficiency), or greater than or equal to (in the case of necessity) the score on the outcome. If the outcome scores are dichotomous (0 or 1), but causal variables are all between zero and one, the researcher may overestimate the extent to which the posited subset relationship exists. Thus, the causal conditions were calibrated to crisp sets to prevent overestimating causal relationships.

the basis for determining which property spaces would be assigned "0" and which would be assigned "1". For each all-cases data set, two tests for sufficiency were run, the first with a consistency cutoff of > 0.00, the second with a consistency cutoff of > 0.1. For the positive-on-outcome approach, consistency scores for sufficiency and coverage scores for necessity were uniformly 1.0 (due to the elimination of all negative cases), and not all property spaces contained cases (limited diversity). For this reason, when editing the truth table, consistency scores were not useful to determine the cutoff; frequency scores were used instead. For each positive-on-outcome data set, the first sufficiency test was run with a frequency cutoff of zero, and the second was run with a frequency cutoff of one. For all sixteen data sets, tests of necessity were also run.

Results

In this section, we present the results of each methodology on the eight simulated data sets.

Logistic Analysis

The logistic analyses coupled with model reduction using likelihood-ratio tests resulted in the "best" models presented in Table 2. For three of the data sets, the full model was retained as the "best" model. For the other five data sets, the model was reduced by between six and eight variables.

[Table 2 about here]

The model-reduction technique did not always reduce the model to include only those variables relevant in the simulation generating the data set. It did sometimes eliminate a variable used in generating the dependent variable. When the full model was retained as the "best" model, the variables with the highest levels of significance were generally (but not always) the variables used in generating the dependent variable for that dataset. A closer comparison is made between these logistic results and the simulated models later in the paper.

Qualitative Comparative Analysis

The results for sufficiency and necessity tests for all generated data sets are in the Table 2 below followed by a more detailed discussion of the models separately.

[Table 3 about here]

Model 1

Sufficiency tests for Model 1 yielded weak results; three of the four tests returned all primitive expressions with low coverage scores. Only one test from the positive-on-outcome data set returned a primitive expression with a modest coverage score of 0.68 (frequency cutoff: 1.00; consistency cutoff: 1.00):

Solution	Raw coverage	Unique coverage	Consistency
$\sim x_3 * x_5$	0.68	0.4	1
$\sim x_1 * \sim x_2 * \sim x_3 * \sim x_4$	0.12	0.04	1
$x_1 * \sim x_2 * x_4 * x_5$	0.14	0.06	1
$\sim x_1 * x_2 * x_4 * x_5$	0.16	0.04	1

solution coverage: 0.82

Tests for necessity, however, yielded more robust results:

Primitive	Consistency (all cases)	Coverage (all cases)	Consistency (positive)	Coverage (positive)
x_1	0.42	0.04	0.42	1
$\sim x_1$	0.58	0.06	0.58	1
x_2	0.42	0.04	0.42	1
$\sim x_2$	0.58	0.06	0.58	1
x_3	0.2	0.02	0.2	1
$\sim x_3^*$	0.8	0.08	0.8	1
x_4	0.6	0.05	0.6	1
$\sim x_4$	0.4	0.04	0.4	1
x_5^*	0.86	0.09	0.86	1
$\sim x_5$	0.14	0.01	0.14	1

Given the nature of the data (rare events), the coverage scores do not yield much information, but the consistency scores indicate the necessity of x3 and x5 for the outcome. Hence, for Model 1:
$\sim x_3 + x_5 \rightarrow Y$

Model 2

Sufficiency tests for Model 2 indicate that $\sim x_2 * \sim x_5$ is weakly sufficient for the outcome. The consistency score is low (0.12), and the coverage score of 0.53 indicates that this causal path explains little more than half the outcome. Necessity tests, on the other hand, point to three variables. As with Model 1, consistency scores were the same across both the all-cases and positive-on-outcome approaches:

Primitive	Consistency (all cases)	Coverage (all cases)
x_1	0.45	0.05
$\sim x_1$	0.55	0.06
x_2	0.28	0.03
$\sim x_2^*$	0.72	0.08
x_3	0.45	0.05
$\sim x_3$	0.55	0.06
x_4^*	0.87	0.09
$\sim x_4$	0.13	0.01
x_5	0.28	0.03
$\sim x_5^*$	0.72	0.07

Consistency scores thus indicate the following relationship of necessity:
$\sim x_2 + x_4 + \sim x_5 \rightarrow Y$

Model 3

As we recall, the underlying relationship of Model 3 is curvilinear. Necessity and sufficiency tests for Model 3 yielded no results. Not until the x_2 variable was recalibrated to reflect a curvilinear relationship[14] do the tests return noteworthy results. The absence of x_2 is shown to be necessary for the outcome, scoring 0.84 for consistency on the necessity tests. This suggests that QCA can detect curvilinear relationships, but the researcher must suspect that the relationship is curvilinear and calibrate the variable accordingly.[15]

Models 4-6

The results from Models 4 show that $x_3 * x_5$ is sufficient, with a consistency score of 0.13 and coverage of 0.63. Individually, variables x_3, x_4, and x_5 appear as necessary causes. Model 5 returns two causal paths for

[14]The x_2 variable was transformed from its original fuzzy values so that low and high values were coded as "0" and medium values were coded as "1". Medium values were considered to fall between 0.25 and 0.75.

[15]To test whether QCA was picking up this relationship accidentally, variables that were known to not be curvilinear in other models were similarly calibrated; no relationship was detected.

the outcome: $x_3* \sim x_4$ and $x_2 * x_3* \sim x_5$, covering 57% and 61% of the outcome set, though not uniquely. Though the coverage score of $x_3* \sim x_4$ is below the coverage threshold determined at the outset of this paper (0.60), it is included here because it is unique coverage (the portion of the outcome it explains on its own) is much higher than that of $x_2 * x_3* \sim x_5$. Unique coverage scores are 0.15 and 0.00, respectively. Individually, each of these variables is detected in tests of necessity as well. Similarly, model 6 returns two causal paths, $x_4* \sim x_5$ and $x_2* \sim x_3 * x_4$, whose variables also feature individually in tests of necessity. The coverage scores for the two sufficient paths are 0.85 and 0.58, respectively. The score for $x_2* \sim x_3 * x_4$ is also below the threshold of 0.60, and is included here because of the high consistency score of 0.30. This is much higher than would be expected with rare events.[16]

Model 7

The test results for model 7 are the weakest, with the $\sim x_2$ and x_5 featuring in tests of necessity, but with very low consistency scores (0.64 and 0.62, respectively). Model 7 proved to be the most difficult model for QCA to detect any underlying relationship.

Model 8

Initially, model 8 showed only $\sim x_5$ to be both necessary and sufficient for the outcome. However, this changed once the x_2 variable was transformed to capture a curvilinear relationship. Before x_2 was recalibrated, tests of necessity and sufficiency indicated it had little to no causal effect on the outcome. After the calibration, x_2 appeared in both tests of necessity, with a consistency score of 0.79, and as part of the causal equation for sufficiency with $\sim x_5$.

Comparing Results

In the table below, we compare the simulated models directly with the logistic and Boolean results. Because Boolean analysis does not have coefficients, the description and results are simplified to indicate positive or negative relationships.[17]

M	Simplified Description	Logit Results (sig. IVs only)	Boolean Results
1	$-x_3$, $+x_5$	$-x_3$, $+x_5$	$-x_3$, $+x_5$
2	$-x_2$, $+x_4$	$+x_4$	$-x_2$, $+x_4$, $-x_5$
3	U-shaped over x_2	U-shaped over x_2	$-x_2$
4	$+x_4$, $+x_3x_5$	$+x_4$, $+x_3x_5$	$+x_3$, $+x_4$, $+x_5$, $+x_3x_5$
5	$+x_2x_3$, $-x_4x_5$	$+x_1$	$+x_2$, $+x_3$, $-x_4$, $-x_5$, $x_3* \sim x_4$, $x_2 * x_3* \sim x_5$
6	$+x_2$, $-x_3$, $+x_4$, $-x_5$	$-x_3$, $+x_4$, $-x_5$	$-x_3$, $+x_4$, $-x_5$, $x_4* \sim x_5$, $x_2* \sim x_3 * x_4$
7	$+x_2$, $+x_5$, $-x_2x_5$	$+x_5$, $-x_2x_5$	$-x_2$, $+x_5$
8	Inverted-U over x_2, $-x_5$	Inverted-U over x_2, $-x_5$	x_2, $-x_5$

Only on model 1 was there agreement on both techniques that matched the simulated data. On models 3 and 8, there was near agreement on both techniques once the Boolean analysis was looking for a curvilinear relationship. This constitutes three positive outcomes for each technique. On the other five models, there were mixed results.

The logistic analysis picked up another positive outcome in arriving at the true description for model 4. The Boolean analysis also uncovered the correct interaction term for model 4 but then made more of the primitive components of that interaction than logit did.

[16] Schneider and Wagemann (2012, pp. 148-149) indicate that there is sometimes a tradeoff between consistency and coverage scores.

[17] For some of the Boolean results, the interactions did not lend themselves to simple positive or negative relationships and were subsequently left as is.

With respect to model 5, the logistic analysis outright fails. In this model, two interaction terms generated the dependent variable. The logistic analysis didn't pick up either interaction term; worse, its one significant result was for a simple independent variable that wasn't even used in generating the dependent variable. However, Boolean analysis did not clearly discern the true relationship either. While Boolean analysis picked up the correct sign on the components of the interaction terms, it did not pick up the true interaction terms themselves.

In model 6, logistic analysis picked up three of the four additive relationships correctly. Boolean analysis picked up the same primitive relationships correctly. However, Boolean analysis then added two interactive causal paths that were not part of the true relationship.

Model 7 represented a traditional interaction term for statistical analysis in the true model. The logistic analysis picked up the correct interaction term and one component, but not the other component. Interestingly, Boolean analysis only picked up the correct primitives but not the interaction term.

Another way to compare the models is by examining their predictive capacity. In Table 4 below, we compare the percentage correctly classified and the positive predictive value for each method's best "solution" for the model dataset. The percentage correctly classified is the percentage of matching cases (i.e., cases $y = 0$ given $\hat{y} = 0$ plus cases with $y = 1$ given $\hat{y} = 1$ divided by the total N). Given the emphasis in Boolean analysis on predicting positive cases, we also examine the "positive predictive value", which is the percentage of cases correctly predicted given those cases where a prediction of a positive outcome was made.

[Table 4 about here]

In this table, we see a common failing of statistical techniques on rare events data. Namely, a very low rate of even making positive predictions. For models 1-3 and 7-8, the logistic analyses do not predict any positive cases; thus, the positive predictive value cannot be calculated. For models 4-6, the positive predictive value of logistic regression is better than that of Boolean analysis. This is largely because logistic analysis is more "tentative" in making positive predictions in the first place. Across all models, Boolean analysis makes more positive predictions that are correct, but does so by making many more positive predictions that are incorrect.

The converse is true for logistic analysis, which has a consistently high percentage correctly classified. This is almost entirely because it is getting all the negative cases predicted correctly. For Boolean analysis, because it over predicts positive cases, it necessarily has a lower percentage correctly classified.

Conclusion

In this brief conclusion, we summarize the main comparisons and contrasting points of our "battling methodologies". In a way, the examined methodologies all have the same starting point. They all have a foundation in set theory, from which "degrees of membership" and "conditional probabilities" are derived. These other concepts are themselves part of probability theory, though QCA authors seem not to attribute this branch of mathematics.

QCA is an inductive method for finding the best-fitting (Boolean algebra) equation given a set of variables while also exploring a large set of interactions among those variables. These best-fitting equations do not provide information regarding the relative importance of variables (i.e., there are no coefficients, standardized or otherwise), but the overall equation combined with the data does provide the conditional probability of correct prediction (within the sample). Statistical methods can be and have been put to the same inductive task, though the focus of training is on theory testing. There are also some examples of theory testing with Boolean algebra.

Measurement is a critical issue for all methodologies. If QCA and/or Boolean algebra were to be a core tool for comparative politics, a best practice would be to think of measurement in terms of the $[0, 1]$ interval. This would avoid the scaling and calibration issue wherein a measure with a larger range needs to be re-scaled into the unit interval before QCA analysis can be conducted.

While some social science concepts are reasonably measured along a "none-to-all" or "absent-to-present" scale, not all are. We do use primary measures that deal with counting things (such as population) or other,

derived measures that are naturally continuous (such as economic wealth). Statistical analysis often has to transform these variables to make the analysis more appropriate (such as log transformation), but it is not a requirement that all such measures be transformed into the unit interval before conducting analysis. This demonstrates a greater flexibility of statistical analysis over QCA, though we would all be better off if we thought more deeply about measurement issues.

Another point of greater flexibility is in the number of cases and the rarity of the dependent variable that logistic analysis can handle compared to QCA/Boolean algebra. One of our first issues was how many cases to have in our simulated datasets. While we chose 1000 somewhat arbitrarily (so that there would be a "medium N" of about 50 positive cases for potential QCA/Boolean analysis), it was clear that QCA/Boolean analysis would not run with the typical tens of thousands cases that statistical analyses of country-years often have. Even with our 1000 cases, the coverage scores for the all-cases Boolean tests were below the threshold that QCA authors would generally be comfortable reporting as meaningful.

Our overall finding was that both methodologies generated similar findings regarding our simulated datasets. In some ways, this is not surprising as both derive their logic from set theory. In fact, QCA seems to be reinventing the wheel of analysis. But by giving existing probability theory concepts new names and conflating some of these concepts with statistical ones, QCA is creating a distinct language that ignores the work of others and makes it harder to converse within the discipline.

Given the greater flexibility of statistical analysis and the fact that QCA seems to be reinventing the wheel, why should we engage in the translation process? There are two potential strengths of QCA worth considering: conjunctural causation and equifinality. Conjunctural causation (or interactions among variables) is the major theme of QCA/Boolean analysis and its "estimator" is designed to calculate coverage for every combination of the independent variables. Equifinality is another potential strength of QCA/Boolean analysis. A typical logit estimation gives a series of *ceteris paribus* statements regarding each of the independent variables. The idea of a "causal path" is only implicit; the idea that there may be multiple "causal paths" is entirely absent. Thus, the notion that a variable might have a positive effect (along one causal path) **and** have a negative effect (along another causal path) is rarely considered.

Table 1. Summary Statistics of the Simulated Models

		x1	x2	x3	x4	x5	y
Model 1	mean	0.4964	0.5004	0.4962	0.5134	0.4921	0.050
	SD	0.2832	0.2845	0.2924	0.2837	0.2869	0.2181
	min	0.0027	0.0004	0.0006	0.0023	0.0023	0
	max	0.9986	1.0000	0.9975	0.9997	0.9992	1
Model 2	mean	0.5079	0.5204	0.5001	0.4990	0.4998	0.053
	SD	0.2891	0.2891	0.2907	0.2942	0.2950	0.2241
	min	0.0003	0.0012	0.0011	0.0005	0.0000	0
	max	0.9990	1.0000	0.9998	0.9980	0.9998	1
Model 3	mean	0.5028	0.5057	0.4866	0.4994	0.5010	0.064
	SD	0.2923	0.2877	0.2909	0.2883	0.2858	0.2449
	min	0.0000	0.0011	0.0008	0.0007	0.0008	0
	max	0.9996	0.9999	0.9996	0.9993	0.9984	1
Model 4	mean	0.4922	0.5037	0.5006	0.4985	0.4889	0.051
	SD	0.2938	0.2856	0.2886	0.2936	0.2875	0.2201
	min	0.0001	0.0005	0.0004	0.0001	0.0006	0
	max	0.9998	0.9991	0.9977	0.9995	0.9980	1
Model 5	mean	0.5124	0.4917	0.5072	0.5044	0.4848	0.049
	SD	0.2855	0.2856	0.2850	0.2882	0.2865	0.2160
	min	0.0019	0.0009	0.0001	0.0000	0.0001	0
	max	1.0000	0.9990	0.9990	0.9981	0.9986	1
Model 6	mean	0.5028	0.5110	0.5001	0.4971	0.4999	0.067
	SD	0.2893	0.2855	0.2887	0.2792	0.2892	0.2501
	min	0.0004	0.0026	0.0012	0.0010	0.0002	0
	max	0.9999	0.9993	0.9999	0.9992	0.9996	1
Model 7	mean	0.5065	0.4763	0.4898	0.5046	0.4934	0.056
	SD	0.2819	0.2938	0.2878	0.2901	0.2837	0.2300
	min	0.0004	0.0011	0.0001	0.0007	0.0006	0
	max	0.9981	0.9993	0.9973	0.9989	0.9996	1
Model 8	mean	0.5089	0.4979	0.5108	0.5007	0.4983	0.056
	SD	0.2903	0.2874	0.2870	0.2868	0.2847	0.2300
	min	0.0013	0.0007	0.0004	0.0004	0.0013	0
	max	0.9997	0.9995	0.9999	0.9987	0.9976	1
Averages	mean	0.5037	0.5009	0.4989	0.5021	0.4948	0.056
	SD	0.2882	0.2874	0.2889	0.2880	0.2874	0.2292
	min	0.0009	0.0011	0.0006	0.0007	0.0007	0
	max	0.9993	0.9995	0.9988	0.9989	0.9988	1

Table 2. Logit Reduced Model Analysis

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
x1		-0.151			**1.561**	0.132		
		0.7681			**0.0092**	0.8237		
x2		-1.387	**-12.904**		7.987	0.632	1.640	**16.026**
		0.5506	**0.0000**		0.0821	0.8280	0.1731	**0.0000**
x3	**-2.499**	-0.393		-0.476	1.501	**-5.404**		
	0.0000	0.7544		0.7279	0.6025	**0.0080**		
x4		**2.690**		**2.383**	-0.291	**6.638**		
		0.0188		**0.0000**	0.7513	**0.0000**		
x5	**4.696**	-2.521		-0.079	0.306	**-11.535**	**3.761**	**-3.084**
	0.0000	0.2272		0.9520	0.9195	**0.0128**	**0.0001**	**0.0000**
x2sq		-0.241	**12.907**		-4.752	2.952		**-16.249**
		0.9080	**0.0000**		0.1040	0.2558		**0.0000**
x5x3		-0.139		**4.071**	1.981	1.946		
		0.9436		**0.0485**	0.5356	0.5815		
x5x4		2.340			-1.230	4.239		
		0.3139			0.5687	0.3612		
x2x3		-1.086			3.326	-1.110		
		0.5650			0.3356	0.7172		
x2x5		0.005			-5.295	2.118	**-4.104**	
		0.9982			0.0573	0.5064	**0.0244**	
_cons	-4.950	-2.892	-0.939	-5.329	-7.852	-4.214	-4.685	-4.653
	0.0000	0.0109	0.0010	0.0000	0.0016	0.0050	0.0000	0.0000
N	1000	1000	1000	1000	1000	1000	1000	1000
r2_p	0.1989	0.1545	0.1213	0.1414	0.2671	0.5375	0.0570	0.1537
chi2	78.9689	64.0563	57.7102	56.9849	104.4559	264.2663	24.5847	66.3605
df_m	2	10	2	4	10	10	3	3
ll	-159.0308	-175.2274	-208.9796	-172.9568	-143.3323	-113.6757	-203.5241	-182.6362

legend: b/p

Independent variables significant at p<0.05 are in bold.

Table 3. Boolean Analysis Results

Model	Solution Formula	Consistency (sufficient)[1]	Coverage (sufficient)	Consistency (necessity)	Coverage (necessity)[2]
Model 1	~x3 + x5 (N)	N/A	N/A	0.80, 0.86	0.08, 0.09
Model 2	~x2+ x4 + ~x5 (N)	N/A	N/A	0.71, 0.86, 0.71	0.08, 0.09, 0.07
Model 3	~x2 (N)	N/A	N/A	0.84	0.11
Model 4	x3 + x4 + x5 (N) x3*x5 (S)	0.13	0.63	0.73, 0.68, 0.78	0.07, 0.07, 0.08
Model 5	x2 + x3 + ~x4 +~x5 (N) x3*~x4 (S) x2*x3*~x5 (S)	0.11, 0.23	0.57, 0.61	0.80, 0.92, 0.61, 0.78	0.08, 0.09, 0.06, 0.07
Model 6	x2+~x3+x4+~x5 (N) x4*~x5 (S) x2*~x3*x4 (S)	0.23, 0.31	0.85, 0.58	0.72, 0.87, 0.96, 0.90	0.09, 0.11, 0.12, 0.12
Model 7	~x2 + x5 (N)	N/A	N/A	0.64, 0.62	0.07, 0.07
Model 8	x2 + ~x5 (N and S)	0.62	0.14	0.79, 0.75	0.09, 0.08

1 Consistency scores reported here are from the all-cases tests.

2 Coverage scores reported here are from the all-cases tests.

Table 4. Comparing Predicted Probabilities

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Logistic Summary								
count(y=1|y_hat_logit=1)	0	0	0	1	4	34	0	0
count(y=0|y_hat_logit=0)	950	947	936	949	946	920	944	944
Positive predictive value	---	---	---	100.0%	44.4%	72.3%	---	---
Correctly classified	95.0%	94.7%	93.6%	95.0%	95.0%	95.4%	94.4%	94.4%
Boolean Summary								
count(y=1|y_hat_boolean=1)	49	53	33	49	49	67	27	49
count(y=0|y_hat_boolean=0)	250	143	488	125	59	59	712	238
Positive predictive value	6.5%	6.2%	6.9%	5.6%	5.2%	7.1%	10.4%	6.5%
Correctly classified	29.9%	19.6%	52.1%	17.4%	10.8%	12.6%	73.9%	28.7%