It's Sets All the Way Down: Algebras, Logic, and the False Foundations of QCA

Abstract

Researchers who use qualitative comparative analysis (QCA) and its variants justify its use by making two related claims. The first is that QCA is, in contrast to quantitative methods, "set-theoretic." The second is that QCA's "set-theoretic" approach allows it to reflect accurately the logic of social science theory. I show that both of these claims are false. First, I demonstrate that the fundamental building blocks of quantitative methods, the random variable and the probability function, are firmly grounded in sets. Moreover, I establish that these quantitative tools are a more flexible and informative way of manipulating sets. Second, I demonstrate that the logic embodied by QCA through Boolean algebra is too weak to capture social scientific theories. Quantitative methods are the real "set-theoretic methods," and they provide a much richer way to understand the social world than QCA.

Kevin A. Clarke[†]

January 25, 2019

[†]Associate Professor, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: kevin.clarke@rochester.edu.

1 Introduction

In the late 1980s, Charles Ragin introduced a new methodology for the social sciences known as qualitative comparative analysis (QCA).[1] The goal, Ragin (1987, x) writes, "is to identify the unique strengths of case-oriented methods and to formalize them as a general method of qualitative comparison using Boolean algebra." Among political scientists, the method has made inroads in all subfields, but the largest group of consumers has been qualitative comparativists. A visit to `compasss.org`, the internet home of QCA, reveals a bibliography of comprising nearly 70 books and hundreds of peer-reviewed articles. Software packages may be downloaded for Stata, R, and UNIX. Aspiring scholars can take QCA training courses around the globe through organizations such as the American Political Science Association, the European Consortium for Political Research, ICPSR, the International Political Science Association, and the Institute for Qualitative and Multi-Method Research.

The growth of QCA remains unabated despite numerous critical assessments (see, for example, Lieberson 2001, 2004; Seawright 2005a,b; Achen 2005; Dunning 2012; Brady 2013; Hug 2013; Braumoeller 2015; Paine 2016; Munck 2016). None of these treatments, however, considers the two related claims that form the foundation of the QCA literature. The first is that QCA's focus on sets differentiates the approach from quantitative methods. Adherents used the adjective "set-theoretic" to distinguish the suite of methods associated with QCA from standard statistical tools. Ragin (2000,

[1]See Appendix A for a brief introduction.

2008), for example, refers to the analysis of "set-theoretic relationships" and claims that correlation methods are incapable of assessing "set-theoretic relations" (Ragin, 2008, 103). Schneider and Wagemann (2012) have written a guide to "set-theoretic" methods. Thiem et al. (2016, 743) use "configurational comparative methods [CCMs]." "Set-theoretic" methods even have acronyms: STM (set-theoretic methods) and STCM (set-theoretic comparative methods).[2]

The second and related claim is that "set-theoretic" methods and, in particular, Boolean algebra, capture the natural language and logic of qualitative comparative political science. Goertz and Mahoney (2012) contrast the foundations of qualitative methods in set theory and logic to the foundations of quantitative methods in probability and statistical theory. They further contrast "the natural language of logic in the qualitative culture with the language of probability and statistics in the quantitative culture" [18]. Thiem et al. (2016, 746) contend that "Boolean algebra establishes the language of CCMs, linear algebra that of RAMs." They go on to claim that the Boolean algebra used by QCA is "incommensurate" with the linear algebra used by quantitative methods.

The distinction between "set-theoretic" methods and the putative "non-set-theoretic" methods strikes political methodologists as odd because sets are the foundation of all statistical practice and inference. Classical statistical methods are "set-theoretic" methods, and they provide a much richer

[2] When referring to quantitative methods, Ragin writes of the "correlational approach" (2000; 2008), while Thiem et al. (2016) use "regressional analytic methods (RAMs)." Schneider and Wagemann (2012, 12) use the term "non-set-theoretic," and many others refer to "large-N analysis."

way to manipulate sets than the so-called "set-theoretic" methods. What truly sets conventional statistics and QCA apart is the latter's embrace of a weak and restrictive form of logic. QCA exists in a nether region between qualitative and quantitative research, and it exhibits neither the depth of qualitative research nor the power of quantitative research.

My discussion unfolds in three parts. In Section 2, I demonstrate that the random variable, a core concept in statistics, is a numerical representation of a set. The same is true of functions of random variables and regression. Quantitative methods can extract far more information from sets than QCA. In Section 3, after a brief introduction to Boolean algebras, I demonstrate the deep connection between Boolean algebras, random variables, and probability. I explain that QCA uses only the simplest kind of Boolean algebra (two-element Boolean algebra). In Section 4, I demonstrate the isomorphism between two-element Boolean algebra and propositional logic and show that propositional logic is too weak to capture social scientific theories.

2 The set-theoretic foundations of statistics

2.1 Preliminaries

In this section, I demonstrate that, contrary to claims by QCA proponents, quantitative methods are firmly grounded in sets. Moreover, the tools available to quantitative researchers allow much richer ways to understand and manipulate sets. The fact that QCA proponents are unaware of the connection between sets and statistics is not altogether surprising; scholars who have not made a concerted study of statistics may fail to realize the deep connection. Econometrics textbooks commonly used by graduate students in political science, e.g. Gujarati and Porter (2008), Wooldridge (2012), and Greene (2011), relegate the discussion to a paragraph in an appendix. Maddala and Lahiri (2010) include a brief paragraph in an early chapter.[3] Students looking for a rigorous discussion of the role of sets need to turn either to a serious probability text, e.g. DeGroot and Schervish (2010) or Casella and Berger (2002), or an intermediate level econometrics text, e.g. Amemiya (1994), Poirier (1995), Spanos (1999), or Bierens (2005).[4] These latter books, however, are often inaccessible to political science students without stronger-than-average backgrounds in mathematics.

I assume that readers are familiar with only the basic operations and notation of set theory including subsets (A is a subset of B, $A \subset B$), unions (A union B, $A \cup B$), intersections (A intersection B, $A \cap B$), inclusion (a is a member of B, $a \in B$), and exclusion (a is not a member of B, $a \notin B$).[5]

To keep the exposition concrete, consider a classic random experiment: spinning a roulette wheel.[6] An American roulette wheel has 38 pockets labelled 1-36 and 0 and 00. 18 of the pockets are red, 18 are black, and the 0 and 00 pockets are green.

The *outcomes set* comprises the possible outcomes of a spin of the wheel

---

[3]To be fair, the authors of these texts assume that students have a had a rigorous course in probability, which is all too often lacking in political science training.

[4]I use Spanos's (1999) notation throughout, and my discussion owes much to his exposition.

[5]This paper cannot substitute for a rigorous course in probability; I suppress some details in the interest of simplicity.

[6]A random experiment is one where all outcomes are known ahead of time; the outcome of any particular trial is unknown, but there is a recognizable pattern associated with the outcomes; and it can be repeated under identical conditions.

Figure 1: An American roulette wheel with two green pockets (0, 00). French roulette wheels have only a single green pocket.

$$S = \{1, 2, \ldots, 36, 0, 00\}.$$

The 1 pocket is an *elementary element* in S, $1 \in S$, and the outcome "landed on red" (R) is an event, which is a subset of the outcomes set, $R \subset S$.[7] The *complement* of R, $\bar{R}$, is the set of outcomes that are in S, but not in R (that is, the outcomes that are black or green).

2.2 Random variables

The world does not produce data; it produces outcomes. When we spin a roulette wheel, we might be interested in the events "landed on red" (R), "landed on black" (B), and "landed on green" (G). We need to assign numbers to these events in order to talk about notions such as expectation

[7] An outcome is an event, but an event is not necessarily an outcome.

and variation. That is the job of a random variable.

Graduate students (and professors) in political science often have difficulty answering the question "What is a random variable?" Answers range from a variable that can take multiple values to a variable that is random. A random variable in actuality is just a function, and like all functions, it is a mapping between two sets. Consider a function that maps from a set A to a set B, $f : A \to B$. The function assigns to every element in the set A a unique element in the set B.

In the case of a random variable, the set of outcomes produced by the world is mapped onto the set of real numbers, $\mathbb{R}$.[8] Denoted by a capital letter, a random variable is the function $X(.) : S \to \mathbb{R}$. We can think of a random variable as a numerical representation of a set.

Let's assume that we are interested in the event "landed on red" when we spin the roulette wheel. The outcomes set is then $S = \{R, B, G\}$. We can represent the event "landed on red" as 1 and "did not land on red" as 0

	$X(.)$	
S	$\to$	$\mathbb{R}$
R		1
B		0
G		0

Note that $X(.)$ is a function; it assigns a unique value in $\mathbb{R}$ to every outcome in S.

Not every function, however, is a random variable. The outcomes of our random experiment, a spin of a roulette wheel, have probabilities associated

[8] Points that can be thought of as lying along an infinite number line.

with them. We know that the probability of the event "landed on red" is the number of possible red pockets divided by the total number of pockets or 18/38. The random variable we defined above assigns the number 1 to the event "landed on red," and we need to ensure that we can assign the probability 18/38 to the number 1. The way we do that is with a *field*, which is also known as an *algebra*.

A collection $\Im$ of subsets of S is a field or algebra if it meets three conditions:

- the outcomes set S, in the the collection, $S \in \Im$;
- if an event A is in the collection, $A \in \Im$, then its complement, $\bar{A}$, must also be in it, $\bar{A} \in \Im$;
- if two events A and B are in the collection, $A, B \in \Im$, then their union is also in it, $A \cup B \in \Im$.

The definition of a field means that $\Im$ is non-empty and is *closed* under complementation, unions, and intersections.[9]

If we demand that our function preserve the field, we can have no problems assigning the correct probabilities. The function we defined above is a random variable relative to the field $\Im = \{S, \emptyset, R, \bar{R}\}$. We can easily check that this collection of subsets meets the three conditions of being an algebra. First, the outcomes set S is in the collection. Second, for every event in the collection, its complement is also included. The complement of S is the empty set, $\emptyset$, and vice versa. The complement of R is $\bar{R}$ and vice versa.

[9] An operation $\circ$ is *closed* if applying $\circ$ to elements in a set A yields an element in A.

Three, the union of any two events must be in the collection. The union of R and $\bar{R}$ is S. The union of S and any other event is S, and the same is true of $\emptyset$. We therefore meet the conditions for a random variable.[10]

2.3 Probability

So far, we have a random variable, X, to translate the set of outcomes, S, and the associated algebra, $\Im$, into the set of real numbers, $\mathbb{R}_X$. We now want to assign probabilities to those numbers that are consist with the probabilities from the roulette wheel. The first step is to define probability which, just like a random variable, is a function that maps sets into numbers. In this case, the function maps from the field, $\Im$, defined above, to the 0-1 interval. The function $\Pr(.) : \Im \to [0, 1]$ is a probability set function if it meets three conditions:

- the probability of the outcomes set is 1, $\Pr(S) = 1$;
- the probability of any event A in the field is greater than or equal to 0, $\Pr(A) \geq 0$, for $A \in \Im$;
- if A_1 and A_2 are mutually exclusive events in $\Im$, the probability of their union is the sum of their probabilities, $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$.

The random variable assigns numbers to subsets of the outcomes set, and the probability function assigns probabilities (values in the 0-1 interval) to those same subsets. We connect these two sets of numbers using a

[10] More technically, we write a random variable as $X(.) : S \to \mathbb{R}_X$, such that $A_x := \{s : X(s) = x\} \in \Im$ for each $x \in \mathbb{R}$, where $\mathbb{R}_X$ is the image of S under X, and the set A_x is the pre-image of X at $X = x$.

density function, $f_x(.)$, which assigns the probabilities from the probability function to the value of the random variable. The density function tells us the probability that the random variable, X, takes on a particular value,

$$f_x(x) := \Pr(X = x) \text{ for all } x \in \mathbb{R}_X.$$[11]

In our example, the probability of "lands on red" is equivalent to $\Pr(X = 1)$, which is 18/38. The probability of "does not land on red" is equivalent to $\Pr(X = 0)$, which is 20/38. The result is a distribution,

x	0	1
$f_x(x)$	18/38	20/38

The probability distribution is a representation of the set of outcomes of the random experiment plus the probabilities associated with those outcomes. Instead of working with the outcomes set, S, the algebra, $\Im$, and the known probabilities, we can now work with numbers that represent the outcomes set, $\mathbb{R}_X$, and the density function, $f_x(x)$.[12] The important thing to keep in mind is that we are still dealing with sets. (See the Appendix for a discussion of countably infinite and infinite outcomes sets.)

2.4 Functions of random variables

Representing sets as random variables allows us to consider operations such as addition, multiplication, or averaging. Here, the difference between QCA

[11] Note that $\Pr(X = x)$ here is shorthand for $A_x := \{s : X(s) = x\}$.

[12] $(\mathbf{S}, \Im, \Pr(.)) \overset{X(.)}{\Rightarrow} (\mathbb{R}_X, f_x(.))$.

and statistics is starkest. When dealing with sets, QCA allows the operations of union, intersection, and complement (see Section 3 for further explanation). However, no measure of central tendency is available; there are no means, medians, or modes. QCA users cannot discuss the notion of spread: there are no variances or ranges. There is no concept of a distribution. Few tools of any kind are available.

In contrast, the tools of quantitative methods make it easy to map from sets to averages. The key to such a mapping is that *a function of random variables is a random variable.*[13] Thus, everything we have previously said about random variables holds for functions of random variables, and functions of random variables are important because *estimators* are functions of random variables.

Consider spinning the roulette wheel twice. All the possible outcomes in terms of color are listed in the outcomes set below

$$S = \{(RR), (RB), (RG), (BB), (BR), (BG), (GG), (GR), (GB)\}.$$

Now, let us define two random variables. $X(.)$ is the numerical representation of getting two red in the two spins. $Y(.)$ is the numerical representation of getting two black in the two spins

[13]Technically, a Borel function of random variables is a random variable.

$$
\begin{aligned}
X(RR) &= 1 \\
X(RB) &= X(RG) = X(BR) = X(GR) = 0 \\
X(BB) &= X(BG) = X(GG) = X(GB) = 0 \\
Y(BB) &= 1 \\
Y(RB) &= Y(BR) = Y(BG) = Y(GB) = 0 \\
Y(RR) &= Y(RG) = Y(GG) = Y(GR) = 0.
\end{aligned}
$$

The joint density of X and Y (see the Appendix for the derivation) is

$x \backslash y$	0	1
0	0.78	0.11
1	0.11	0.00

We can determine the density of a random variable Z that is the sum of X and Y, $Z = X + Y$. Recall that X can take on the values $\mathbb{R}_X = \{0, 1\}$, and Y can take on the values $\mathbb{R}_Y = \{0, 1\}$. When we add the possible values, we see that Z can take on the values $\mathbb{R}_Z = \{0, 1, 2\}$. We can find the distribution of Z by noting that, for example, $Z = 0$ only when $X = 0$ and $Y = 0$, which is 0.78 of the time. $Z = 1$ when $X = 0$ and $Y = 0$ or when $X = 1$ and $Y = 0$. Thus, $\Pr(Z = 1) = 0.11 + 0.11 = 0.22$. $Z = 2$ only occurs when both random variables equal 1. The density of Z is therefore

z	0	1	2
$f_z(z)$	0.78	0.22	0

Z is a random variable like any other random variable. That is, it is a mapping from a set of outcomes to numbers in such a way that preserves the event structure (field). In the case of Z, the mapping is

S	$\overset{Z(.)}{\rightarrow}$	$\mathbb{R}$
1 red and 1 black		0
2 red or 2 black		1
2 red and 2 black		2

The density function for Z expresses these outcomes along with their associated probabilities.

Functions of random variables are important because estimators are functions of random variables, $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$. Thus, an estimator is also a random variable in exactly the same way as all the other random variables we have discussed. The sample mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

for example, is an estimator, which is a random variable. We can derive probabilities attached to the outcomes, which would give us a density function known as a *sampling distribution.* None of these tools are available to QCA users.

2.5 Regression

Finally, regression analysis is also based on sets. First, the regression estimator, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is, just like the estimators above, a function of random variables and therefore a random variable. We also know that regression provides the mean of a dependent variable Y conditional on an independent variable(s) X, $E[Y|X]$, and we can view this conditional expectation as a random variable (Spanos, 1999).

Consider again our joint density along with the marginal densities (summing across the rows and columns)

$x \backslash y$	0	1	$f(x)$
0	0.78	0.11	0.89
1	0.11	0.00	0.11
$f(y)$	0.89	0.11	1.00

The *conditional density* of Y at the value $X = 0$ is[14]

y	0	1
$f(y\|x=0)$	0.876	0.124

and the conditional density of Y at the value $X = 1$ is

y	0	1
$f(y\|x=1)$	1.00	0.00

[14] 0.78/0.88 and 0.11/0.88

The conditional expectations are therefore

$$\begin{aligned} E[Y|X=0] &= (0)*0.876+(1)*0.11=0.11 \\ E[Y|X=1] &= (0)*0.11+(1)*0.0=0 \end{aligned}$$

The conditional expectation is therefore a random variable

x	0	1
$\Pr(X=x)$	0.89	0.11
$E[Y\|X]$	0.11	0

Spanos (1999, 359) points out that the conditional means can be viewed in terms of a random variable Z

$$Z(.) : E[Y(.)|\sigma(X(.))] : S \to \mathbb{R},$$

where $\sigma(X(.))$ is a different notation for an algebra or field.[15]

2.6 Discussion 1

The term "set-theoretic method" is meant to distinguish the suite of procedures associated with QCA from conventional statistical analysis. In fact, proponents of QCA and its variants use the term "set-theoretic" as a selling point. Ragin (2008, 13), for example, argues that "all social science theory is verbal and, as such, is formulated in terms of sets and set relations." In

[15] A σ-algebra is an extension of algebras to outcomes that are countably infinite.

response to criticism, Ragin (2005, 37) writes, "QCA is based on the algebra of sets, not on linear algebra, the basis of regression analysis. QCA's analytic engine is fueled by *set-theoretic relations*, not correlations." I have shown that conventional statistic analysis is deeply rooted in sets, and thus any benefits accruing from being "set-theoretic" must also apply to quantitative tools. More importantly, QCA users have only a limited menu of tools with which to manipulate and describe sets. Quantitative methods have no such restrictions.

QCA proponents may object that quantitative methods may be practiced without consciousness of the set-theoretic foundations of the tools being used. Running a regression does not require explicit consideration of sets. Of course, neither does QCA. In Appendix A, I list the six stages of a QCA analysis. A researcher can perform each stage without worrying about unions, intersections, or complements. The only truly set-theoretic piece of QCA is the use of Boolean algebra, the subject of the next section.

3 Boolean algebras and statistics

Claims that "set-theoretic" methods stand apart from conventional statistical analysis generally reference QCA's use of Boolean algebra. Schneider and Wagemann (2012, 8), for example, write that QCA makes use of truth tables and the "principles of logic minimization." They define logical minimization as "applying the rules of Boolean algebra" [329]. Proponents of QCA then claim that QCA and statistics are incommensurate because the former uses Boolean algebra while the latter uses linear algebra, and the laws governing the respective algebras are different. Thiem et al. (2016, 748), for example, points out that in Boolean algebra, $x \vee (-x) = 1$, while in linear algebra, $x + (-x) = 0$.[16] Of course, we can make similar demonstrations regarding the differences between scalar algebra and matrix algebra. In scalar algebra, $a * b = 0$ leads to the conclusion that either a or b is 0. In matrix algebra, the fact that $\mathbf{AB} = \mathbf{0}$ does not imply that either $\mathbf{A}$ or $\mathbf{B}$ is 0. No one, however, considers scalar and matrix algebra to be incommensurate.

We need to know more about Boolean algebras to understand where the intersection with classical statistics lies. As I detail below, random variables and probability are defined on Boolean algebras.

3.1 Preliminaries

When people hear of Boolean algebra, they immediately think of sets comprising 0s and 1s, but that is just the simplest example of a (nondegenerate) Boolean algebra. A Boolean algebra is a non-empty set S, with two binary operations *meet* ($\vee$) and *join* ($\wedge$), a *complement* ($\prime$), two distinguished elements 0 and 1, and that satisfies the following axioms[17]:

[16] Thiem et al. (2016) uses + for $\vee$, but I use $\vee$ to avoid confusion later in this paper.
[17] Other laws can be derived from these laws.

Identity laws

$p \wedge 1 = p$ $\qquad$ $p \vee 0 = p$

Complement laws

$p \wedge p' = 0$ $\qquad$ $p \vee p' = 1$

Commutative laws

$p \wedge q = q \wedge p$ $\qquad$ $p \vee q = q \vee p$

Distributive laws

$p \wedge (q \vee r) = (p \wedge q) \vee (p \wedge r)$ $\qquad$ $p \wedge (q \vee r) = (p \vee q) \wedge (p \vee r)$

Parts of this definition require additional elucidation. A binary operation on a set is a rule that assigns to each ordered pair of elements of the set some element of the set.[18] So if we use $\vee$ and $\wedge$ on any subsets of a set, the result is also a member of that set. The complement or negation is a unary operator, which means that it is a rule that assigns an element of the set some element of the set.[19] If we use $'$ on any subset of a set, the result is also a member of that set. Finally, a distinguished element is an identity element for a binary operation. That is, we need an element e such that $p \wedge e = p$ and $p \vee e = p$, and 0 and 1 are those two elements.

Nothing that we have said so far demands that a Boolean algebra includes only 0s and 1s. All we need to form a Boolean algebra are two binary operations, a unary operation, and two elements that satisfy the identity requirements (and meet the axioms). Consider, for instance, the set $X = \{1, 2, 3, 6\}$. X along with the following operations

- $a \vee b = lcm(a, b)$, $\forall\, a, b \in X$,

[18] More technically, a binary operation on a set A is a function from $A \times A$ into A.

[19] More technically, a unary operation on a set A is a function from A into A.

- $a \wedge b = gcd(a, b), \forall\, a, b \in X$,

- $a' = 6/a, \forall\, a, b \in X$

is a Boolean algebra. The binary operations are lcm (lowest common multiple) and gcd (greatest common divisor). The unary operation is $6/a$. What are the distinguished elements? For the "zero" element, we need $p \vee 0 = p$ and $p \wedge p' = 0$. 1 is the "zero" element because $p \vee 1 = lcm(p, 1) = p$, and $p \wedge p' = gcd(p, 6/p) = 1$. For the "one" element, we need $p \wedge 1 = p$ and $p \vee p' = 1$. 6 is the "one" element because $p \wedge 6 = gcd(p, 6) = p$ and $p \vee p' = lcm(p, 6/p) = 6$.

A straightforward way to form a Boolean algebra is through the power set $\mathcal{P}(\cdot)$ of a set S. Consider the set $S = \{R, B, G\}$ from Section 2.2. The power set of S, $\mathcal{P}(S)$, is the set of all subsets of S, or

$$\mathcal{P}(S) = \{\emptyset, (R), (B), (G), (RB), (RG), (BG), S\}.$$

$\mathcal{P}(S)$ is a Boolean algebra with the following translations:

$\emptyset$	$\rightarrow$	0
Union ($\cup$)	$\rightarrow$	Join ($\vee$)
Intersection ($\cap$)	$\rightarrow$	Meet ($\wedge$)
Not ($\bar{A}$)	$\rightarrow$	Complement (A')
S	$\rightarrow$	1

Although a full proof that the power set is a Boolean algebra is beyond the scope of this paper, we can see the connection by substituting any one of the subsets of $\mathcal{P}(S)$ for p, q, and r in the Boolean laws above. Consider the subset (R):

$$R \cap S = R \text{ becomes } R \wedge 1 = R, \text{ and}$$
$$R \cup \bar{R} = S \text{ becomes } R \vee R' = 1.$$

It should be clear that $\mathcal{P}(S)$ properly translated is a Boolean algebra.

A Boolean algebra where S is empty is a degenerate Boolean algebra, and the simplest non-degenerate Boolean algebra comprises just two elements, 0 and 1.

I should note here, as it will become important later in the paper, that some Boolean algebras have special properties, and the two-element Boolean algebra mentioned above is one of those algebras. For example, the properties

$$x \vee y = 1 \text{ if and only if } x = 1 \text{ or } y = 1,$$
$$x \wedge y = 0 \text{ if and only if } x = 0 \text{ or } y = 0,$$

only hold for two-element Boolean algebras.

3.2 Boolean algebras, fields, and probability

Note that $\mathcal{P}(S)$ is closed under the set-theoretic operations of union, intersection, and complementation just as fields are (see Section 2.2). In fact, fields are Boolean algebras. Every field that I discuss in the first part of this paper is a Boolean algebra. Consider, for example, the field that I define in Section 2.2: $\Im = \{S, \emptyset, R, \bar{R}\}$. $\Im$ is a Boolean algebra if we make the same translation that we did in the case of the power set above (and the power set of any finite outcomes set is also a field). Thus, there is a deep connection between random variables and Boolean algebras.

QCA scholars have little choice but concede that random variables are based on Boolean algebras, but once the sets are mapped to the real line and become numbers, they may well claim that the "setness" disappears. That is, the numbers do not have to obey the laws of Boolean algebra although the sets upon which they are based do. For example, even when the random variable is binary, $x + (-x) \neq 1$, although in terms of sets $x \vee (-x) = 1$. It is now time to bring probability back into the discussion. Recall that the reason for defining a random variable on a field is to ensure that probabilities are assigned that do not violate the "setness" of the outcomes.

Consider a random variable X with density:

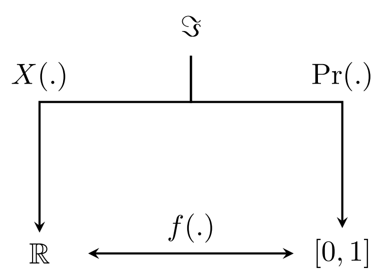
x	0	1	2	3
$f_x(x)$	0.25	0.30	0.25	0.20

It is certainly true that if we sum 0 and "not 0," we get 6, not 1. However, if we add the probability of 0 and the probability of " not 0," we do get 1,

$$\Pr(0) + \Pr(\bar{0}) = 0.25 + 0.75 = 1.$$

By the same token, it must be the case that $p \wedge p' = 0$. Well, the probability of 0 and "not 0" occurring together is 0. These results suggest that there is a connection between probability and Boolean algebra and, in fact, probability is defined as a function on a Boolean algebra.[20] This particular fact is fairly esoteric, but the point is that all aspects of classical statistics are based on sets.

Figure 2: The relationship between fields, random variables, and probability. The functions $X(.)$ and $\Pr(.)$ map the real line $\mathbb{R}$ and the 0-1 interval, respectively, while preserving the field, $\Im$. The density function $f(.)$ connects them.



Consider Figure 2. A random variable $X(.)$ maps subsets of a set (a Boolean algebra) to a subset of the real line while preserving the field. Probability $\Pr(.)$ maps subsets of a set (a Boolean algebra) to the [0,1] interval while preserving the field. Finally, the density function $f(.)$ connects the random variable to the [0,1] interval. At no point in the structure depicted in Figure 2 does the "setness" of the field get lost.

3.3 Discussion 2

I have only scratched the surface of Boolean algebras, but it is clear that there is a deep connection between Boolean algebras and random variables, between Boolean algebras and probability, and between random variables

[20]Probability is "a norm $p : \mathcal{B} \rightarrow [0, 1]$ on a Boolean algebra $\mathcal{B}$ of events" (Primas, 1999, 587). Similarly, "*the theory or probability studies a collection of objects, which form a normed Boolean algebra*; these objects are called *events* and the norm $p(A)$ of an event A is called its *probability*" (Yaglom and Yaglom, 1973, 42).

and probability. The claim that QCA and quantitative methods are distinct because Boolean algebra and linear algebra obey different laws is both superficial and largely irrelevant. The foundation of statistical analysis is not built on "probability and statistical theory," but on sets; probability and statistical theory flow from them.

Why QCA uses the two-element Boolean algebra, and the limitations that imposes, is the subject of the next section.

4 Boolean algebras and logic

Any reader familiar with QCA may wonder, after learning about Boolean algebras, why QCA (the original variant) is limited to the two-element Boolean algebra, and the answer has to do with the connection between the two-element Boolean algebra and logic. Although researchers writing about "set-theoretic" methods often conflate the two, logic and Boolean algebra are not the same thing. Two-element Boolean algebra is a model for a kind of logic that is too weak to capture all but the simplest logical reasoning. It cannot capture temporal order, quantification (e.g., "for all" or "there exists" statements), or other forms of qualitative argumentations such as generalization.

4.1 Preliminaries

There are many different types of logics including propositional logic, syllogistic logic, predicate logic, modal logic, as well as non-classical logics such as intuitionistic or constructive logic. QCA is solely concerned with propositional (also known as classical or sentential) logic, and we shall see, only some Boolean algebras model propositional logic.

A proposition (or sentence) is a statement that is either true or false, but not both. Commands and questions are not propositions, and neither are statements such as $X \geq 5$ because whether the statement is true depends on the value of X. Examples of propositions include:

- David Collier studies American politics;
- The United States of America is a democracy;
- The Red Sox will win the 2019 World Series.

Whether true or not, all three statements are either true or false. Propositional logic concerns these kinds of simple sentences and their connectives (and, or, not). The sentences or propositions are denoted by symbols such as P, Q, and S, and the user defines what they mean (the semantics). So I could assign the proposition "David Collier studies American politics" to the letter P. These propositional symbols are known as atoms or atomic formulas because they contain no connectives. Thus, the proposition represented by P is the smallest unit that the logic can analyze. Molecules or molecular formulas are formed out of atoms and the connectives (and, or, not). If I assign the letter Q to the proposition "The United States of America is a democracy," then "P and Q" ($P \wedge Q$) is a molecular formula.

A well-formed formula in propositional logic obeys the following rules:

- P is a well-formed formula;

- If P is a well-formed formula, then so is $\neg P$;
- If P and Q are well-formed formula, then so are $P \wedge Q$, $P \vee Q$, $P \to Q$, and $P \leftrightarrow Q$.

Propositional logic is truth-functional, which means that the truth of a molecular formula is a function of the truth values of the atomic propositions in it. That is, every proposition is either true or false, and every connective is truth-functional. We can therefore assess the truth value of a formula such as $(P \vee Q) \wedge S \to P$ by considering the truth values of the atoms P, Q, and S. Truth functions are also known as Boolean functions.

The two-element Boolean algebra is a model of propositional logic because we can interpret "true" as 1, "false" as 0, and the connectives (and, or, not) as meet, join, and complement. To see the connection more completely, remember that two-element Boolean algebras have special properties. These are,

$$x \vee y = 1 \text{ if and only if } x = 1 \text{ or } y = 1,$$

$$x \wedge y = 0 \text{ if and only if } x = 0 \text{ or } y = 0,$$

We can write these special properties in what is known as an arithmetic table for the join operator $\vee$ and the meet operator $\wedge$,

$\vee$	0	1
0	0	1
1	1	1

$\wedge$	0	1
0	0	0
1	0	1

We can compare these arithmetic tables to the truth tables for the logical "or" and "and" in propositional logic,

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

In the arithmetic table for $\vee$, whenever a 1 appears on top or the left side, the outcome of the join operator is 1. In the truth table for "or", whenever a T appears under P or Q, a T appears under $P \vee Q$. The operations take on the value 0 or F only when both inputs are 0 or F. In the arithmetic table for $\wedge$, only when a 1 appears on top and the left side is the outcome 1. In the truth table for "and", only when T appears under P and Q does a T appears under $P \wedge Q$. The operations take on the value 1 or T only when both inputs are 1 or T. The correspondences are one-to-one.

A less simple Boolean algebra, such as the power set $\mathcal{P}(S)$ discussed above, is not a model of propositional logic. We can interpret S as "true" and $\emptyset$ as "false," but what of the other values in the set? QCA is inextricably tied to propositional logic because it uses two-element Boolean algebra.

4.2 The limitations of propositional logic

Propositional logic, although of immense importance to computer design (0s and 1s) and electrical engineering (on and off), has serious limitations, and contrary to the claims of QCA proponents, social theory is often too complex to be rendered in propositional logic. Consider, for example, the following perfectly sound argument:

All democracies have a legislature.

The United States is a democracy.

The United States has a legislature.

How could we express this argument in propositional logic? Recall that the smallest unit of analysis, the atom, is a true or false sentence, and we assign a unique letter to each sentence. The logic ignores the structure of the atom. That is, propositional logic cannot "see" inside the sentences (Jeffrey, 1991). In the second premise, there is an object (the United States) and a property of that object (a legislature). Both, however, are hidden when we assign a letter. So we might assign L to the first premise, D to the second premise, and U to the conclusion. The result, however, $L, D \vdash U$ is not valid (if the premises are true, the conclusion must be true). U does not necessarily follow from $L\,, D$. We can see this from the left truth table in Figure 3.

In the second line of table, we see that both premises are true, but the conclusion is false. The takeaway is not that the argument is false, but that argument cannot be proven in propositional logic. The only valid rendering of the argument requires making a different argument. If the premises were "If the United States is a democracy, then the United States has a legislature" ($D \rightarrow L$) and "the United States is a democracy" (D), we could validly conclude that "the United States has a legislature L, or

Figure 3: Two propositional logics. The logic on the left is invalid because the conclusion is not necessarily true when both premise are true. The logic on the right is valid because the conclusion is true when both premises are true.

L	D	U
T	T	T
T	T	F
T	F	T
F	T	T
T	F	F
F	T	F
F	F	T
F	F	F

$D \to L$	D	L
T	T	T
F	T	F
T	F	T
F	F	F

$D \to L, D \vdash L$. The validity of this logic is demonstrated in the right truth table in Figure 3. When both premises are true, the conclusion is true.

We cannot use propositional logic to identity traits or properties that units may have in common such as "democracies have legislatures." That, in turn, makes it difficult, if not impossible, to generalize or talk of patterns (Lemmon, 1992). The implications for QCA are dramatic. Goertz and Mahoney (2012, 193), for example, being their discussion of "qualitative generalizations" by pointing to statements that "All A are B." As we have seen, however, such statements cannot be represented in propositional logic.

The weaknesses of propositional logic led to the development of predicate or first-order logic, which introduced quantifiers (the universal $\forall$ and the existential $\exists$) , variables, and relations. "All democracies have a legislature" is easily translated as $\forall x(democracy(x) \to legislature(x))$. Two-element Boolean algebra, however, is not a model for predicate logic, which requires

Halmos algebra or cylindric algebra.[21]

4.3 Discussion 3

A core claim made by QCA proponents is that qualitative scholars speak the language of logic, and that QCA captures that language. The problem with this claim is that the logic represented by the two-element Boolean algebra, propositional logic, is a weak language. That leaves two possibilities: either the language used by qualitative scholars is simple enough to be represented by the two-element Boolean algebra, or QCA is not adequately capturing that language. Some combination of these two possibilities gets us close to the truth. I do not mean to imply that qualitative research is simplistic, far from it. Qualitative scholars, like quantitative scholars, have, however, struggled with a limited causal language (that of necessity and sufficiency) with its roots in the 18th century. The result is that nuance is sacrificed in favor of something like quantitive generalizations. The list of "All A are B" generalization provided by Goertz and Mahoney (2012, 194) contains a number of "almosts" (as in "almost all") and exceptions. At the same time, many well-established results in social science cannot be captured by QCA. Arrow's Theorem, which relies heavily on quantifiers, cannot be written in propositional logic. The problem is not the use of logic; it is the use of limited and restrictive logic.

[21] A cylindric algebra is a Boolean algebra with an additional unary operation and an additional distinguished element that capture existential and universal quantification.

5 Conclusion

QCA practitioners believe in two core claims regarding their methodology. The first is that their method is "set theoretic," which sets it apart from conventional quantitative techniques. The second is that logic is the language of qualitative scholars, and that language is well represented by Boolean algebra. Neither of these claims is true. Quantitative methods are grounded entirely in sets. The building blocks of statistics, the random variable and probability, preserve the nature of the sets that they represent. Classical statistical methods are "set-theoretic" methods, and they allow a much more diverse set of tools for describing and manipulating sets. QCA is based on propositional logic, which is a weak logical language. It is too weak , for example, to allow generalizations or easily recognized forms of argumentation.

In effect, QCA occupies an unnecessary middle ground between qualitative research and quantitative research. It has neither the depth nor the nuance of the best qualitative work, while simultaneously failing to provide either the power or scope of the best quantitative work. QCA's focus on sets is not unique, and the logic that it employs is unable to represent contemporary social theory (whether formal or not). The worst-case scenario for those who value qualitative research is to find themselves inextricably tied to a methodology that has so little to offer. QCA is holding qualitative research back, not moving it forward.

Appendices

A Introduction to qualitative comparative analysis (QCA)

A full introduction to QCA is well beyond the scope of this paper, and the interested reader should turn to Ragin (1987), Goertz and Mahoney (2012), or Schneider and Wagemann (2012). Seawright (2005a,b) provides a critical introduction. The broad outlines of the method, however, are easy to grasp.

Rihoux and Ragin (2009) list six stages of a QCA analysis:

1. Build the data table;
2. Construct a truth table using conditions drawn from cases;
3. Resolve contradictory configurations;
4. Reduce the truth table using Boolean minimization;
5. Consider 'logical remainders' cases;
6. Interpret the results.

The keys steps to understanding QCA are steps 2, 4, and 6. A researcher creates a truth table by first creating a data matrix comprising cases as the rows, variables (or conditions in QCA parlance) as the columns, and an outcome variable (condition) in the right-most column. The values in the data matrix are then dichotomized (1 for present, 0 for absent), and redundant configurations (combinations of 0s and 1s associated with an outcome) are removed. The result is a truth table, such as the ones in Section 4 where 1s replace Ts and 0s replace Fs.

In step 2, the researcher uses Boolean minimization to reduce the table. That is, the researcher attempts to remove logically redundant conditions in order to merge rows and find a simplified Boolean expression for the outcome. Consider the following truth table for two conditions (P and Q) and outcome (R):

P	Q	R
F	F	F
F	T	F
T	F	T
T	T	T

R takes that value T whenever P takes the value T. Thus, we can write down a Boolean expression for R that can be reduced to P:

$$R = P(\neg Q) + PQ = P(\neg Q + Q) = P(1) = P.$$

When a truth table can be reduced no further, and the result is a combination of conditions, the result is known in QCA as a prime implicant.[22]

The final step is interpretation. As Brown (1990) notes, the theory of prime implicants was developed by Quine and Blake with two different applications in mind. Quine's interest was in finding simplified formulas, and Blake's interest was inference (drawing conclusions from the truth table).

[22]Technically, a prime implicant of a Boolean function f is an implicant of f(a term that implies f) that would no longer be an implicant of f if one of its literals (a letter or complement) were removed. That is, a prime implicant is minimal.

This tension seemingly remains in the practice of QCA. Early practitioners, such as its inventor, Charles Ragin, thought more in terms of formula simplification, while noting the connection between Boolean algebra and necessity and sufficiency. Later practitioners, such as Schneider and Wagemann (2012), focus much more on finding necessary and sufficient conditions.

B Countably infinite and infinite outcomes sets

The outcomes generated by the roulette example in Section 2 are countable and finite. If the outcomes are uncountable, then we cannot use the density function to assign probabilities because the probability that a random variable would take a particular value, $\Pr(X = x)$, is zero. Age is a easy example of such a variable, and we would write the set as

$$S = \{x : x \in \mathbb{R}, 0 < x < \infty\},$$

which indicates that age can take any real number greater than 0. The probability that anyone is 36 years, 10 days, 4 hours, 3 minutes, and 34 seconds old (we could go on here) is 0. We therefore need a new way of organizing our random variable and assigning probabilities.

When dealing with uncountable outcomes, we use intervals, specifically, the half-open interval, $(-\infty, x]$. It turns out that the set of half-open intervals generates an algebra (known as a Borel field). The random variable preserves the structure of the field just as it does with countable outcomes (which is a special case).[23]

[23] $X(.) : S \to \mathbb{R}$, such that $A_x \coloneqq \{s : X(s) \leq x\} \coloneqq X^{-1}\{(-\infty, x]\} \in \Im$ for all $x \in \mathbb{R}$.

To assign probabilities to this kind of random variable, we use the *cumulative distribution function* (CDF), $F_X(.)$, instead of the density function. The CDF is a function that maps from the real numbers to the 0-1 interval (probabilities must be between 0 and 1),

$$F_X(.) : \mathbb{R} \to [0, 1],$$

where $F_X(x) = \Pr((-\infty, x])$. Instead of giving the probability that the random variable will take a particular value, $\Pr(X = x)$, the CDF gives the probability that the random variable will take a value less than or equal to a particular value, $\Pr(X \leq x)$. We can then get the density function by taking the derivative of the CDF with respect to x. Countable or uncountable, a probability distribution that comprises a random variable and associated probabilities is a representation of a set. It is all about sets.

C Multiple random variables

In actual research, we rarely deal with a single random variable. Consider spinning the roulette wheel twice. All the possible outcomes in terms of color are listed in the outcomes set below

$$S = \{(RR), (RB), (RG), (BB), (BR), (BG), (GG), (GR), (GB)\}.$$

Now, let us define two random variables. $X(.)$ is the numerical representation of getting two red in the two spins. $Y(.)$ is the numerical representation of getting two black in the two spins

$$\begin{aligned}
X(RR) &= 1 \\
X(RB) &= X(RG) = X(BR) = X(GR) = 0 \\
X(BB) &= X(BG) = X(GG) = X(GB) = 0 \\
Y(BB) &= 1 \\
Y(RB) &= Y(BR) = Y(BG) = Y(GB) = 0 \\
Y(RR) &= Y(RG) = Y(GG) = Y(GR) = 0.
\end{aligned}$$

Just as we did before, we can represent these sets of outcomes with their probabilities in a probability distribution

x	0	1
$f_x(x)$	0.776	0.224

y	0	1
$f_y(y)$	0.776	0.224

The random variable X maps from sets to the real line, $X(.) : S \to \mathbb{R}$, and so does Y, $Y(.) : S \to \mathbb{R}$. The setup is flexible enough, however, that we can consider these random variables jointly. Let $Z(.)$ be a mapping between the sets of outcomes into two set of numbers, $Z(.) \coloneqq (X(.), Y(.)) : S \to \mathbb{R}^2$. Instead of the events $(X = x)$ and $(Y = y)$, we are looking at their intersection $(X = x) \cap (Y = y)$.

We determine the *joint density function* by writing down the possible outcomes and attaching probabilities

$$
\begin{aligned}
(X=0,Y=0) &= \{(RB),(RG)(BR),(BG), & f(x=0,y=0) &= 0.78 \\
&\quad (GG),(GR),(GB)\} \\
(X=0,Y=1) &= \{(BB)\}, & f(x=0,y=1) &= 0.11 \\
(X=1,Y=0) &= \{(RR)\}, & f(x=1,y=0) &= 0.11 \\
(X=1,Y=1) &= \emptyset, & f(x=1,y=1) &= 0
\end{aligned}
$$

The joint density is in more compact notation is

$x\backslash y$	0	1
0	0.78	0.11
1	0.11	0.00

The joint density is a numerical representation of two sets and the dependency between them. From this point, we can define *conditional densities* and *marginal densities*, which in turn allows us to define a *random sample*.[24]

[24] A random sample is a set of identical and independently distributed (IID) random variables.

References

Achen, C. H. (2005, Spring). Two cheers for charles ragin. *Studies in Comparative International Development 40*(1), 27–32.

Amemiya, T. (1994). *Introduction to Statistics and Econometrics.* Cambridge, MA: Harvard University Press.

Bierens, H. J. (2005). *Introduction to the Mathematical and Statistical Foundations of Econometrics.* NY: Cambridge University Press.

Brady, H. E. (2013, February). Do two research cultures imply two scientific paradigms? *Comparative Political Studies 46*(2), 252–265.

Braumoeller, B. F. (2015, Autumn). Guarding against false positives in qualitative comparative analysis. *Political Analysis 23*(4), 471–487.

Brown, F. M. (1990). *Boolean Reasoning: The Logic of Boolean Equations.* Boston, MA: Kluwer.

Casella, G. and R. L. Berger (2002). *Statistical Inference* (2 ed.). Pacific Grove, CA: Duxbury.

DeGroot, M. H. and M. J. Schervish (2010). *Probability and Statistics* (4 ed.). Reading, MA: Addison-Wesley.

Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach.* NY: Cambridge University Press.

Goertz, G. and J. Mahoney (2012). *A Tale of Two Cultures: Qualitative and*

Quantitative Research in the Social Sciences. Princeton, NJ: Princeton University Press.

Greene, W. H. (2011). *Econometric Analysis* (7 ed.). NY: Pearson.

Gujarati, D. N. and D. C. Porter (2008). *Basic Econometrics* (5 ed.). NY: McGraw-Hill.

Hug, S. (2013, Spring). Qualitative comparative analysis: How inductive use and measurement error lead to problematic inference. *Political Analysis 21*(2), 252–265.

Jeffrey, R. (1991). *Formal Logic: Its Scope and Limits* (3 ed.). NY: McGraw-Hill.

Lemmon, E. (1992). *Beginning Logic*. Indianapolis, IN: Hackett.

Lieberson, S. (2001). Review essay on fuzzy set social science, by charles ragin. *Contemporary Sociology 30*(4), 331–334.

Lieberson, S. (2004, Fall). Comments on the use and utility of qca. *Qualitative Methods 2*(2), 13–14.

Maddala, G. and K. Lahiri (2010). *Introduction to Econometrics* (4 ed.). NY: Wiley.

Munck, G. L. (2016, May). Assessing set-theoretic comparative methods: A tool for qualitative comparativists? *Comparative Political Studies 49*(6), 775–780.

Paine, J. (2016, May). Set-theoretic comparative methods: Less distinctive than claimed. *Comparative Political Studies 49*(6), 703–741.

Poirier, D. J. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: The MIT Press.

Primas, H. (1999). Basic elements and problems of probability theory. *Journal of Scientific Exploration 13*(4), 579–613.

Ragin, C. C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press.

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

Ragin, C. C. (2005, Spring). Core versus tangential assumptions in comparative research. *Studies in Comparative International Development 40*(1), 33–38.

Ragin, C. C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago, IL: The University of Chicago Press.

Rihoux, B. and C. C. Ragin (2009). *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. London: Sage.

Schneider, C. Q. and C. Wagemann (2012). *Set-Theoretic Methods for the Social Sciences: A Guid to Qualitative Comparative Analysis*. Cambridge, NY: Cambridge University Press.

Seawright, J. (2005a, Spring). Assumptions, causal inference, and the goals of qca. *Studies in Comparative International Development 40*(1), 39–42.

Seawright, J. (2005b, Spring). Qualitative comparative analysis vis-a-vis regression. *Studies in Comparative International Development 40*(1), 3–26.

Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data.* NY: Cambridge University Press.

Thiem, A., M. Baumgartner, and D. Bol (2016, May). Still lost in translation! a correction of three misunderstandings between configurational comparativists and regressional analysts. *Comparative Political Studies 49*(6), 742–774.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach* (5 ed.). Mason, OH: Thomson South-Western.

Yaglom, A. and I. Yaglom (1973). *Probability and Information.* Dordrecht, NL: D. Reidel Publishing Company.