The European Political Science Association, 2018 *doi:10.1017/psrm.2018.45*

Causal Modeling with Multi-Value and Fuzzy-Set Coincidence Analysis [*]

MICHAEL BAUMGARTNER AND MATHIAS AMBÜHL

Coincidence Analysis *(CNA) is a configurational comparative method of causal data analysis that is related to* Qualitative Comparative Analysis *(QCA) but, contrary to the latter, is custom-built for analyzing causal structures with multiple outcomes. So far, however, CNA has only been capable of processing dichotomous variables, which greatly limited its scope of applicability. This paper generalizes CNA for multi-value variables as well as continuous variables whose values are interpreted as membership scores in fuzzy sets. This generalization comes with a major adaptation of CNA's algorithmic protocol, which, in an extended series of benchmark tests, is shown to give CNA an edge over QCA not only with respect to multi-outcome structures but also with respect to the analysis of non-ideal data stemming from single-outcome structures. The inferential power of multi-value and fuzzy-set CNA is made available to end users in the newest version of the R package* ***cna****.*

Since the mid-1980ies, different variants of *configurational comparative methods* (CCMs) have gradually been added to the toolkit for causal data analysis in the social sciences. CCMs are designed to investigate different hypotheses and uncover different properties of causal structures than traditional regression analytical methods (RAMs) and, thus, complement the latter (rather than compete with them). RAMs examine covariation hypotheses as "the more/less of X, the more/less of Y" that link *variables*, and they quantify net-effects and effect sizes. CCMs, by contrast, study implication hypotheses as "$X{=}\chi_i$ is (non-redundantly) sufficient/necessary for $Y{=}\gamma_i$" that link *specific values* of variables. Moreover, instead of quantifying effect sizes, CCMs place a Boolean ordering on sets of causes by locating their elements on the same or different causal paths to the ultimate

[*] We thank Alrik Thiem, Tim Haesebrouck, Simon Hug, and two anonymous PSRM referees for very helpful comments on previous versions of this article. Moreover, we have profited from discussions with audiences at the 5th International QCA Expert Workshop, Zürich 2017, and at the ECPR Joint Sessions Workshop "Configurational Thinking in Political Science: Theory, Methodology, and Empirical Application", Nottingham 2017. Finally, we are grateful for generous support by the Swiss National Science Foundation (grant no. PP00P1_144736/1) and by the Bergen Research Foundation and University of Bergen (grant no. 811886).

outcome. In other words, while RAMs investigate the quantitative properties of causal structures as characterized by statistical or probabilistic theories of causation (e.g. Suppes 1970), CCMs scrutinize their Boolean properties as described by regularity theories of causation (Mackie 1974).

The Boolean properties of causation encompass three complexity dimensions. The first is *conjunctivity*: to bring about an effect, say, liberal democracy in early modern Europe (D=1), different factors need to be instantiated (or *not* instantiated) jointly; for instance, according to Downing's (1992) theory of the origins of liberal democracy, a country must have a history of medieval constitutionalism (C=1) and absent military revolutions (R=0). Only a coincident instantiation of the conjunction C=1$*R$=0 produces the effect D=1. *Disjunctivity* is a second complexity dimension: an effect can be brought about along alternative causal paths. Downing (1992, 78-79, 240) identifies four paths leading to the absence of military revolution (R=0): a geography that deters invading armies (G=1), commercial wealth (W=1), foreign resource mobilization (M=1), and foreign alliances (A=1). Each condition in the disjunction G=1 + W=1 + M=1 + A=1 can bring about the effect R=0 independently of the other conditions. The third complexity dimension is *sequentiality*: effects tend to cause further effects, propagating causal influence along causal chains. In Downing's theory there are multiple chains, for instance, W=1 is causally relevant to R=0, which, in turn, is causally relevant to D=1, or there is a chain from A=1 via R=0 to D=1. Overall, the theory entails the following Boolean causal model (cf. Goertz 2006, 254), where "→" stands for the Boolean operation of implication:

$$(G{=}1 \;+\; W{=}1 \;+\; M{=}1 \;+\; A{=}1 \;\rightarrow\; R{=}0) * (C{=}1{*}R{=}0 \;\rightarrow\; D{=}1) \tag{1}$$

The most prominent CCM is *Qualitative Comparative Analysis* (QCA; Ragin 2008). While the original variant of QCA introduced in Ragin (1987), *crisp-set* QCA (csQCA), is restricted to modeling dichotomous variables, there meanwhile exist fully worked out variants that can process multi-value variables, *multi-value* QCA (mvQCA) (Cronqvist and Berg-Schlosser 2009), and variables with continuous values from the unit interval, *fuzzy-set* QCA (fsQCA) (Ragin 2009). However, all QCA variants focus on the complexity dimensions of conjunctivity and disjunctivity only, as QCA treats exactly one factor as endogenous and all other analyzed factors as exogenous. QCA will thus not find a chain model as (1).[1]

In light of this restriction, Baumgartner (2009) introduced a new configurational comparative method called *Coincidence Analysis* (CNA). As a member of the family of CCMs, CNA—just like QCA—investigates implication hypotheses and scrutinizes the Boolean properties of causation. Contrary to QCA, however, CNA is capable of analyzing

[1] In a recent comment on CNA, Thiem (2015) argues that QCA is not necessarily tied to an algorithm that is restricted to single-outcome structures. Thiem then suggests a QCA approach to searching for chains that much resembles CNA.

multi-outcome structures and, hence, of uncovering all Boolean complexity dimensions. CNA is tailor-made to recover chain models as (1).

So far, though, CNA has only been available in a crisp-set variant (csCNA). This paper removes that limitation by generalizing the method for multi-value variables (mvCNA) and variables with continuous values from the unit interval that are interpreted as membership scores in fuzzy sets (fsCNA). This generalization comes with a major adaptation of the basic algorithmic protocol on the basis of which CNA builds causal models. In a nutshell, while CNA so far—just like QCA—adopted a *top-down approach* to model building that first identifies complete sufficient and necessary conditions of outcomes and then gradually eliminates redundant elements, the generalized variant of CNA uses a *bottom-up approach* that progressively combines factor values to complex but redundancy-free sufficient and necessary conditions.

The CNA algorithm presented here has been implemented in a new version of the R package **cna**, version 2.1.1 (Ambühl and Baumgartner 2018), which makes the whole inferential power of mvCNA and fsCNA available to end-users. By drawing on this software package and the currently most reliable R package for QCA, **QCApro** (Thiem 2018),[2] the paper also performs a whole battery of benchmark tests that evaluate and compare the performance of CNA and QCA when applied to data with varying forms of data deficiencies. The test series reveals that the reversal of the basic model building approach gives CNA an edge over QCA not only with respect to multi-outcome structures but also with respect to the analysis of non-ideal data stemming from single-outcome structures.

The paper is organized as follows. We first introduce the theoretical background of CNA along with its main input parameters. Next, the generalization of the CNA algorithm is presented. The final section then reports the results of the test series evaluating and comparing CNA and QCA. The online appendix contains supplemental information on configurational homogeneity and correctness and provides charts with all numeric results of our test series. A replication script detailing all analytical steps is available in the PSRM dataverse.

Theoretical Background

Boolean Difference-making

As all CCMs, CNA searches for causal dependencies as defined by so-called *regularity theories* of causation, whose development dates back to David Hume (1999 (1748)).

[2] Recently, version 3.2 of the **QCA** R package (Duşa 2007) has been published, which introduces very promising new functionalities. However, the default parameter settings of that package are still such that, too often, many data-fitting models are not recovered.

Modern regularity theories define causation in terms of Boolean difference-making within a fixed causal background. More specifically, $X{=}\chi_i$ is a regularity theoretic cause of $Y{=}\gamma_i$ if there exists a (fixed) configuration of background conditions $\mathcal{F}$ such that, in $\mathcal{F}$, a change from $X{=}\chi_k$ to $X{=}\chi_i$, where $\chi_i \neq \chi_k$, is systematically and non-redundantly associated with a change from $Y{=}\gamma_k$ to $Y{=}\gamma_i$, where $\gamma_i \neq \gamma_k$. If $X{=}\chi_i$ does not make a difference to $Y{=}\gamma_i$ in any context $\mathcal{F}$, $X{=}\chi_i$ is redundant to account for $Y{=}\gamma_i$ and, thus, no cause of $Y{=}\gamma_i$ (Mackie 1974; Graßhoff and May 2001; Baumgartner 2013).

To render that idea more precise, some conceptual preliminaries are required. Regularity theoretic causation holds between *variables/factors taking on specific values*. (We will use the terms "variable" and "factor" interchangeably.) Factors represent categorical properties that partition sets of units of observation (cases) either into two sets, in case of binary properties, or into more than two (but finitely many) sets, in case of multi-value properties. Factors representing binary properties can be *crisp-set* (cs) or *fuzzy-set* (fs); the former can take on 0 and 1 as possible values, whereas the latter can take on any (continuous) values from the unit interval. Factors representing multi-value properties are called *multi-value* (mv) factors; they can take on any of an open (but finite) number of possible values $\{0, 1, 2, \ldots, n\}$. Values of a cs or fs factor X are interpretable as membership scores in the set of cases exhibiting the property represented by X. As is conventional in Boolean algebra, we shall abbreviate membership in a set by upper case and non-membership by lower case Roman letters; that is, we write "X" for $X{=}1$ and "x" for $X{=}0$. An alternative interpretation, which lends itself particularly well for causal modeling, is that "X" stands for the presence of the factor X and "x" for its absence. In case of mv factors, we will not abbreviate value assignments and, instead, use the explicit 'Variable=value' notation by writing, say, "$X{=}3$" for X taking the value 3.

Apart from the Boolean operations of conjunction, disjunction, and negation, whose classical definitions are presupposed here, the implication operator "$\rightarrow$" and the equivalence operator "$\leftrightarrow$" are of core importance for the regularity theoretic definition of causation. According to a classical interpretation, an expression as "$X{=}3 \;\rightarrow\; Y{=}4$" states that whenever X takes the value 3, Y takes 4; or "$X \;\rightarrow\; Y$" states that whenever X is present, Y is present. These claims are true if, and only if (iff), there is no case satisfying the left-hand side of "$\rightarrow$" and not satisfying the right-hand side. Furthermore, "$X{=}3 \;\leftrightarrow\; Y{=}4$" and "$X \;\leftrightarrow\; Y$" are true iff the implication holds both ways, meaning that all cases satisfying the left-hand side of "$\leftrightarrow$" also satisfy the right-hand side, and vice versa.

For the subsequent generalization of CNA for fs factors the classical Boolean operations must be translated into fuzzy logic. There exist numerous systems of fuzzy logic (for an overview cf. Hájek 1998), each of which comes with its own rendering of Boolean operations. We will adopt the following fuzzy-logic renderings, which have become standard in the context of CCMs: conjunction $X{*}Y$ is defined in terms of the minimum membership score in X and Y, i.e. $\min(X, Y)$, disjunction $X + Y$ in terms of the maximum membership score in X and Y, i.e. $\max(X, Y)$, negation $\neg X$ (or x) in terms of $1 - X$, an implication $X \rightarrow Y$ is taken to express that the membership score in X is smaller or equal to Y ($X \leq Y$), and an equivalence $X \leftrightarrow Y$ that the membership scores in X and Y are equal ($X = Y$).

Based on the implication operator the notions of *sufficiency* and *necessity* are defined, which are the two Boolean dependencies exploited by regularity theories: X is *sufficient* for Y iff $X \rightarrow Y$ holds; and X is *necessary* for Y iff $Y \rightarrow X$ holds. Analogously, the more complex expression X=3 + Z=2 is sufficient and necessary for Y=4 iff X=3 + Z=2 $\leftrightarrow$ Y=4 holds.

Boolean dependencies of sufficiency and necessity amount to mere patterns of co-occurrence of factor values; as such, they carry no causal connotations whatsoever. In fact, most Boolean dependencies do not reflect causal dependencies. For that reason, regularity theories rely on a *non-redundancy principle* as an additional constraint to filter out those relations of sufficiency and necessity that are due to underlying causal dependencies: A Boolean dependency structure is causally interpretable only if it does not contain any redundant elements. Causes are those elements of sufficient and necessary conditions for which at least one configuration of background conditions $\mathcal{F}$ exists in which they are indispensable to account for a scrutinized outcome. In other words, whatever can be removed from sufficient and necessary conditions without affecting the latter's sufficiency and necessity is redundant and, therefore, not causally interpretable. Only sufficient and necessary conditions that are completely free of redundant elements, viz. minimal, possibly reflect causation (Baumgartner 2015).

Boolean Causal Models

Modern regularity theories formally cash this idea out on the basis of the notion of a minimal theory. Its complete definition is intricate and beyond the scope of this paper (for the latest definition see Baumgartner and Falk 2018). For our subsequent purposes, the following rough characterization will suffice. There are atomic and complex minimal theories. An *atomic minimal theory* of an outcome Y is a minimally necessary disjunction of minimally sufficient conditions of Y. A conjunction Φ of coincidently instantiated factor values (e.g. $X_1{*}X_2{*}\ldots{*}X_n$) is a minimally sufficient condition of Y iff Φ is sufficient for Y ($\Phi \rightarrow Y$), and there does not exist a proper part Φ' of Φ such that $\Phi' \rightarrow Y$. A proper part Φ' of Φ is the result of eliminating one or more conjuncts from Φ. A disjunction Ψ of minimally sufficient conditions (e.g. $\Phi_1 + \Phi_2 + \ldots + \Phi_n$) is a minimally necessary condition of Y iff Ψ is necessary for Y ($Y \rightarrow \Psi$), and there does not exist a proper part Ψ' of Ψ such that $Y \rightarrow \Psi'$. A proper part Ψ' of Ψ is the result of eliminating one or more disjuncts from Ψ. Overall, an atomic minimal theory of Y states an equivalence of the form $\Psi \leftrightarrow Y$ (where Ψ is an expression in disjunctive normal form[3] and Y is a single

[3] A Boolean expression is said to be in disjunctive normal form iff it is a disjunction of one or more conjunctions of one or more literals.

factor value). Atomic minimal theories can be conjunctively concatenated to *complex minimal theories*.

Minimal theories connect Boolean dependencies, which—by themselves—are purely functional and non-causal, to causal dependencies: those, and only those, Boolean dependencies that appear in minimal theories can stem from underlying causal dependencies. Atomic minimal theories stand for causal structures with one outcome, complex theories represent multi-outcome structures. To further clarify the causal interpretation of minimal theories, consider the following complex exemplar:

$$(A{*}b \;+\; a{*}B \;\leftrightarrow\; C) * (C{*}f \;+\; D \;\leftrightarrow\; E) \tag{2}$$

Functionally put, (2) claims that the presence of A in conjunction with the absence of B (i.e. b) as well as a in conjunction with B are two alternative minimally sufficient conditions of C, and that $C{*}f$ and D are two alternative minimally sufficient conditions of E. Moreover, both $A{*}b + a{*}B$ and $C{*}f + D$ are claimed to be minimally necessary for C and E, respectively. Against the background of a regularity theory, these functional relations entail the following causal claims: (i) the factor values listed on the left-hand sides of "↔" are directly causally relevant for the factor values on the right-hand sides; (ii) A and b are located on the same causal path to C, which differs from the path on which a and B are located, and C and f are located on the same path to E, which differs from D's path; (iii) $A{*}b$ and $a{*}B$ are two alternative indirect causes of E whose influence is mediated on a causal chain via C. More generally put, minimal theories ascribe causal relevance to their constitutive factor values, place them on the same or different paths to the outcomes, and distinguish between direct and indirect causal relevancies. That is, they render transparent the three Boolean complexity dimensions of causality—which is why we shall likewise refer to minimal theories as *Boolean causal models*.

Two fundamentals of the interpretation of Boolean causal models must be emphasized. First, ordinary Boolean models make claims about causal relevance *but not about causal irrelevance*. With some additional constraints that are immaterial for our current purposes (for details see Baumgartner 2013), a regularity theory defines X_1 to be a cause of an outcome Y iff there exists a fixed configuration of context factors $\mathcal{F} = X_2{*}\ldots{*}X_n$ in which X_1 makes a difference to Y—meaning that $X_1{*}\mathcal{F}$ and $x_1{*}\mathcal{F}$ are systematically associated with different Y-values. While establishing causal relevance merely requires demonstrating the existence of at least one such difference-making context, establishing causal irrelevance would require demonstrating the non-existence of such a context, which is impossible on the basis of the non-exhaustive data samples that are typically analyzed in observational studies. Correspondingly, the fact that, say, G does not appear in (2) does not imply G to be causally irrelevant to either C or E. The non-inclusion of G simply means that the data from which model (2) has been derived do not contain evidence for the relevance of G.

Second, Boolean models are to be interpreted relative to the data set δ from which they have been derived. They do not purport to reveal all of an underlying causal structure's Boolean properties but only detail those causally relevant factor values along with those conjunctive, disjunctive, and sequential groupings for which δ contains evidence. By extension, two different Boolean models $\mathbf{m}_i$ and $\mathbf{m}_j$ derived from two different data sets δ_i and δ_j are in no disagreement if the causal claims entailed by $\mathbf{m}_i$ and $\mathbf{m}_j$ stand in a subset relation. For example, model (3) does not conflict with model (2):

$$(A \ + \ B \ \leftrightarrow \ C) * (C \ + \ D \ \leftrightarrow \ E) \tag{3}$$

(3) identifies A and B as alternative direct causes of C and indirect causes of E, moreover C and D are claimed to be alternative direct causes of E. All of this also follows from (2). The causal claims entailed by (3) thus constitute a subset of the claims entailed by (2). The two models describe properties of one and the same underlying causal structure at different degrees of detail and relative to different data $\delta_{(2)}$ and $\delta_{(3)}$.

Data, Consistency, Coverage

CCMs analyze *configurational data* δ that have the form of $m \times k$ matrices, where m is the number of units of observation (cases) and k is the number of factors. We subsequently refer to the set of factors $\mathbf{F}$ in an analyzed δ as the *factor frame* of the analysis. While QCA requires that $\mathbf{F}$ be partitioned—prior to the analysis—into a first subset $\{Y\}$ comprising exactly one endogenous factor and a second subset $\mathbf{F} \setminus \{Y\}$ comprising all exogenous factors of the analysis, CNA can dispense with such a partition. If prior causal knowledge is available as to what factors in $\mathbf{F}$ are possible effects and what factors can be excluded as effects, this information can be given to CNA via an optional argument called a *causal ordering*. A causal ordering is a relation $X_i \prec X_j$ defined on the elements of $\mathbf{F}$ entailing that X_j cannot be a cause of X_i (e.g. because X_i is instantiated temporally before X_j). If an ordering is provided, CNA only searches for Boolean models in accordance with the ordering; if no ordering is provided, CNA treats all values of the factors in $\mathbf{F}$ as potential outcomes and explores whether a causal model for them can be inferred from δ.

As real-life data tend to feature noise induced by unmeasured causes of endogenous factors, strictly sufficient or necessary conditions for an outcome Y often do not exist. To still extract some causal information from such data, Ragin (2006) has imported *consistency* and *coverage* measures (with values from the interval $[0, 1]$) into the QCA protocol. Both of these measures are also serviceable for the purposes of CNA. Informally put, consistency (con) reproduces the degree to which the behavior of an outcome obeys a corresponding sufficiency or necessity relationship or a whole model, whereas coverage (cov) reproduces the degree to which a sufficiency or necessity relationship or a whole model accounts for the behavior of the corresponding outcome (for formal definitions see the vignette of the **cna** package, Ambühl and Baumgartner 2018, §3.2). If no (strictly Boolean) relations of sufficiency and necessity with $con = 1$ and $cov = 1$ can be inferred from δ, CNA invites its users to lower the consistency and coverage thresholds con_t and cov_t. For example, by lowering con_t to 0.8, CNA is given permission to treat X as sufficient for Y, even though in 20% of the cases X is not associated with Y. Or by lowering cov_t to 0.8, CNA is allowed to treat X as necessary for Y, even if 20% of the cases featuring Y do not feature X.

Lowering con_t and cov_t must be done with great caution, for the lower these thresholds, the higher the chance that causal fallacies are committed. In QCA, however, it is common to only impose lowest bounds—e.g. 0.75—for the consistency of configurations comprising all exogenous factors, so-called *minterms*. This approach does not guarantee that the consistencies of issued minimally sufficient conditions (or *prime implicants*, as they are called in QCA) and of resulting Boolean models are also above the chosen threshold. Accordingly, the models output by QCA often do not meet the consistency threshold set by the user (cf. the replication script for examples). Moreover, it is common QCA practice not to require lowest bounds for coverage. In consequence, QCA models frequently cover less than half of the cases featuring the outcome in δ.

In CNA, the consistency and coverage standards are higher—for two reasons. First, the sufficient conditions that are ultimately causally interpreted by CCMs are not minterms (which are mere intermediate calculation devices for QCA) but redundancy-free conditions contained in Boolean models. Hence, consistency thresholds must be imposed on the latter, not on the former. Second, a model's coverage being low means that it only accounts for few instances of an outcome in δ. Or differently, in many cases in δ where the outcome is present there are causes at work that are not contained (i.e. unmeasured) in the factor frame $\mathbf{F}$. However, unmeasured causes tend to confound δ—in particular, when they are associated with both exogenous and endogenous factors in $\mathbf{F}$. The presence of confounders casts doubts on the causal interpretability of all dependencies manifest in δ, for uncontrolled causes might be covertly responsible for them. That is, the more likely it is that the data are confounded, the less reliable a causal interpretation of resulting models becomes. The higher the coverage, the less likely it is that we are facing data confounding, the more reliable a causal interpretation of issued models becomes. The online appendix A provides an extended discussion of the conditions under which CNA can be expected to output correct models.

Generalizing the CNA Algorithm

Top-down vs. Bottom-up Search

The goal of CCMs is to infer Boolean causal models from configurational data. The previous section has shown that Boolean functions are amenable to a causal interpretation only if they identify redundancy-free sufficient and necessary conditions, and thus amount to minimal theories that reach imposed consistency (con_t) and coverage (cov_t) thresholds.

There exist two different strategies for building minimal theories: they can be built from the *top down* or from the *bottom up*. The top-down approach proceeds as follows. First, complete sufficient minterms are identified that meet con_t; second, elements are eliminated as redundant as long as the remaining conditions continue to satisfy con_t; third, the minimally sufficient conditions are disjunctively combined to necessary conditions that meet cov_t; fourth, elements are eliminated as redundant that are not required to satisfy cov_t. By contrast, the bottom-up approach starts with single factor values and tests whether they meet con_t; if that is not the case, it proceeds to test conjunctions of two factor values, then to conjunctions of three, and so on. Whenever a conjunction meets con_t (and no proper part of it has previously been identified to meet con_t), it is automatically redundancy-free, that is, a minimally sufficient condition (msc), and supersets of it do not need to be tested for sufficiency any more. Then, the bottom-up approach tests whether single msc meet cov_t; if not, it proceeds to disjunctions of two, then to disjunctions of three, and so on. Whenever a disjunction meets cov_t (and no proper part of it has previously been identified to meet cov_t), it is automatically redundancy-free, viz. a minimally necessary condition, and supersets of it do not need to be tested for necessity.

Both QCA and the original variant of csCNA adopt versions of the top-down approach—albeit in very different algorithmic implementations (cf. Baumgartner 2015). By contrast, the generalization of CNA developed here reverses the direction of model building. Prima facie, it might seem that it does not matter whether models are built from the top down or from the bottom up because both directions should ultimately lead to the same results. Although that is indeed the case for some data types, in particular for ideal data, it does not hold generally. For instance, when applied to data that do not allow for modeling outcomes with perfect consistency, it can happen that—contrary to the bottom-up approach—the top-down approach does not succeed in eliminating all redundancies from sufficient conditions. The reason is that when building models from the top down it is (implicitly) presumed that consistency threshold violations are *monotonic* in the following sense: if a factor C cannot be eliminated from a sufficient condition $A{*}B{*}C$ because $A{*}B$ alone does not meet con_t, then C *plus* some further factor from $A{*}B$ cannot be eliminated either. Therefore, if eliminating C from $A{*}B{*}C$ leads to a violation of con_t, the top-down approach concludes that C is needed to account for the outcome, meaning C is a difference-maker. That conclusion, however, is not valid, because consistency threshold violations are not monotonic.

To see this, consider the data matrix in Table 1a, for which the following consistencies hold:

$$\begin{aligned} con(A{*}B{*}C \;\rightarrow\; D) &= {}^{3}\!/_{4} = 0.75 \\ con(A{*}B \rightarrow D) &= {}^{8}\!/_{11} = 0.73 \\ con(A \rightarrow D) &= {}^{15}\!/_{20} = 0.75 \end{aligned}$$

That is, if con_t is set to 0.75, the condition $A{*}B{*}C$, which is sufficient for D with $con = 0.75$, satisfies the threshold. By contrast, $A{*}B$, which results from $A{*}B{*}C$ by eliminating C, falls short of con_t. Nonetheless, further eliminating B lifts the remaining condition above con_t again, as A alone is sufficient for D with $con = 0.75$. That means, while C initially

A	B	C	D	n
1	1	1	1	3
1	1	1	0	1
1	1	0	1	5
1	1	0	0	2
1	0	1	0	1
1	0	0	1	7
0	1	1	0	4
1	0	0	0	1
0	0	0	0	3

(a)

A	B	C	D
1.00	1.00	0.60	1.00
1.00	1.00	0.40	1.00
0.40	1.00	0.40	0.10
1.00	0.30	0.30	0.00

(b)

A	B	C	OUT_D	n	con
1	1	1	0	1	0.647
1	1	0	0	1	0.647
0	1	0	0	1	0.167
1	0	0	0	1	0.000

(c)

Table 1

Note: Table a is a *cs* data matrix, where the right-most column lists the number of cases featuring a configuration and D is the outcome. Table b is an fs data matrix with outcome D, and c the corresponding QCA truth table at $con_t = 0.75$.

appears to be a non-redundant element of the sufficient condition $A{*}B{*}C$, it turns out to be redundant after all. The top-down approach, however, only tests the removability of single factors at a time and infers that a condition is redundancy-free if removing single factors would push that condition below con_t. Therefore, at $con_t = 0.75$, a procedure that adopts the top-down approach as QCA issues model (4) for Table 1a.

$$A{*}b{*}c \; + \; A{*}B{*}C \longrightarrow D \qquad con = 0.83;\; cov = 0.67 \tag{4}$$

In contrast, by first testing whether single factors meet con_t, the bottom-up approach directly finds that A itself is sufficient for D. Moreover, it turns out that A is necessary for D, as it accounts for D with perfect coverage. Overall, at $con_t = 0.75$, a procedure that builds models from the bottom up issues model (5).

$$A \leftrightarrow D \qquad con = 0.75;\; cov = 1 \tag{5}$$

(5) is preferable to (4), for two reasons. First, the product of consistency and coverage, which is a common measure for overall model fit, is significantly higher for (5). Second, model (5) only ascribes causal relevance to A, whereas (4) also determines B, C and their negations to be causes of D, even though the data in Table 1a do not contain evidence that these factors actually make a difference to D at $con_t = 0.75$. Hence, when applied to noisy data, the top-down approach runs a risk of drawing causal inferences that go beyond the data.

Also, the top-down approach may abandon an analysis prematurely. To see this, consider the fs data in Table 1b, where D is the outcome.[4] The four configurations

[4] We thank Tim Haesebrouck for this example.

	A	B	C	D
c_1	0	0	0	0
c_2	0	1	0	0
c_3	1	1	0	0
c_4	0	0	1	0
c_5	1	0	0	1
c_6	1	0	1	1
c_7	0	1	1	1
c_8	1	1	1	1

(a) *cs data*

	A	B	C	D
c_1	1	3	3	1
c_2	2	2	1	2
c_3	2	1	2	2
c_4	2	2	2	2
c_5	3	3	3	2
c_6	2	4	3	2
c_7	1	3	3	3
c_8	1	4	3	3

(b) *mv data*

	A	B	C	D	E
c_1	0.17	0.02	0.15	0.26	0.09
c_2	0.97	0.23	0.73	0.08	0.10
c_3	0.10	0.72	0.61	0.38	0.08
c_4	0.64	0.73	0.82	0.12	0.66
c_5	0.11	0.30	0.06	0.99	0.78
c_6	0.69	0.23	0.91	0.98	0.84
c_7	0.31	0.80	0.62	0.65	0.74
c_8	0.65	0.87	0.92	0.82	0.85

(c) *fs data*

TABLE 2 *Data types analyzed by CNA.*

in that data have the consistencies listed in the last column of the QCA truth table in Table 1c, meaning that, if con_t is set to 0.75, none of the configurations are sufficient for the outcome. A top-down procedure as QCA abandons the analysis at this point. That, however, is unwarranted because there in fact exists a Boolean model for D that meets $con_t = 0.75$ and moreover reaches perfect coverage:

$$A{*}B \leftrightarrow D \qquad con = 0.78;\ cov = 1 \tag{6}$$

By starting the analysis with single factors, the bottom-up approach finds model (6) in the second iteration.

To avoid the problems of the top-down approach, the generalization of CNA developed in this paper builds models from the bottom up.

The Essentials of the CNA Algorithm

The generalized CNA algorithm takes as mandatory inputs (i) a data set δ, (ii) con_t and cov_t thresholds, and (iii) an upper bound called $maxstep$ for the maximal complexity of atomic solution formulas (atomic causal models) to be built. $Maxstep$ serves the pragmatic purpose of keeping the search space computationally tractable in reasonable time. The user can set it to any complexity level if computational time is not an issue. Optionally, CNA can be given a causal ordering.

Contrary to QCA, which first transforms the data into an intermediate calculative device called a *truth table*, the CNA algorithm operates directly on the data. Data processed by CNA can either be of type "crisp-set" (cs), "multi-value" (mv) or "fuzzy-set" (fs). Examples of each data type are given in Table 2. In what follows, we first discuss the generalized CNA algorithm in the abstract, using the explicit 'Variable=value' notation, and then we illustrate its procedural steps on the basis of the fs data in Table 2c.

CNA causally models configurational data δ over a factor frame $\mathbf{F}$ in four stages:

Stage 1 On the basis of a provided ordering, CNA first builds a set of potential outcomes

$\mathbf{O} = \{O_h{=}\omega_f, \ldots, O_m{=}\omega_g\}$ from the factor frame $\mathbf{F} = \{O_1, \ldots, O_n\}$ in δ, where $1 \leq h \leq m \leq n$, and second assigns a set of potential cause factors $\mathbf{C}_{O_i}$ from $\mathbf{F} \setminus \{O_i\}$ to every element $O_i{=}\omega_k$ of $\mathbf{O}$. If no ordering is provided, all value assignments to all elements of $\mathbf{F}$ are treated as possible outcomes in case of *mv* data, whereas in case of *cs* and *fs* data $\mathbf{O}$ is set equal to $\{O_1{=}1, \ldots, O_n{=}1\}$.

Stage 2 CNA attempts to build a set $\textsc{msc}_{O_i=\omega_k}$ of minimally sufficient conditions that meet con_t for each $O_i{=}\omega_k \in \mathbf{O}$. To this end, it first checks for each value assignment $X_h{=}\chi_j$ of each element of $\mathbf{C}_{O_i}$, such that $X_h{=}\chi_j$ has a membership score above 0.5 in at least one case in δ, whether the consistency of $X_h{=}\chi_j \longrightarrow O_i{=}\omega_k$ in δ meets con_t, i.e. whether $con(X_h{=}\chi_j \longrightarrow O_i{=}\omega_k) \geq con_t$. If, and only if, that is the case, CNA puts $X_h{=}\chi_j$ into the set $\textsc{msc}_{O_i=\omega_k}$. Next, CNA checks for each conjunction of two factor values $X_m{=}\chi_j * X_n{=}\chi_l$ from $\mathbf{C}_{O_i}$, such that $X_m{=}\chi_j * X_n{=}\chi_l$ has a membership score above 0.5 in at least one case in δ and no part of $X_m{=}\chi_j * X_n{=}\chi_l$ is already contained in $\textsc{msc}_{O_i=\omega_k}$, whether $con(X_m{=}\chi_j * X_n{=}\chi_l \longrightarrow O_i{=}\omega_k) \geq con_t$. If, and only if, that is the case, CNA puts $X_m{=}\chi_j * X_n{=}\chi_l$ into the set $\textsc{msc}_{O_i=\omega_k}$. Next, conjunctions of three factor values with no parts already contained in $\textsc{msc}_{O_i=\omega_k}$ are tested, then conjunctions of four factor values, etc., until either all logically possible conjunctions of the elements of $\mathbf{C}_{O_i}$ have been tested or *maxstep* is reached. Every non-empty $\textsc{msc}_{O_i=\omega_k}$ is passed on to the third stage.

Stage 3 CNA attempts to build a set $\textsc{asf}_{O_i=\omega_k}$ of atomic solution formulas (atomic causal models) for every $O_i{=}\omega_k \in \mathbf{O}$, which has a non-empty $\textsc{msc}_{O_i=\omega_k}$, by disjunctively concatenating the elements of $\textsc{msc}_{O_i=\omega_k}$ to minimally necessary conditions of $O_i{=}\omega_k$ that meet cov_t. To this end, it first checks for each single condition $\Phi_h \in \textsc{msc}_{O_i=\omega_k}$ whether $cov(\Phi_h \longrightarrow O_i{=}\omega_k) \geq cov_t$. If, and only if, that is the case, CNA puts Φ_h into the set $\textsc{asf}_{O_i=\omega_k}$. Next, CNA checks for each disjunction of two conditions $\Phi_m + \Phi_n$ from $\textsc{msc}_{O_i=\omega_k}$, such that no part of $\Phi_m + \Phi_n$ is already contained in $\textsc{asf}_{O_i=\omega_k}$, whether $cov(\Phi_m + \Phi_n \longrightarrow O_i{=}\omega_k) \geq cov_t$. If, and only if, that is the case, CNA puts $\Phi_m + \Phi_n$ into the set $\textsc{asf}_{O_i=\omega_k}$. Next, disjunctions of three conditions from $\textsc{msc}_{O_i=\omega_k}$ with no parts already contained in $\textsc{asf}_{O_i=\omega_k}$ are tested, then disjunctions of four conditions, etc., until either all logically possible disjunctions of the elements of $\textsc{msc}_{O_i=\omega_k}$ have been tested or *maxstep* is reached. Every non-empty $\textsc{asf}_{O_i=\omega_k}$ is passed on to the fourth stage.

Stage 4 CNA attempts to build a set $\textsc{csf}_{\mathbf{O}}$ of complex solution formulas (complex causal models) encompassing all elements of $\mathbf{O}$. To this end, CNA conjunctively combines exactly one element from every non-empty $\textsc{asf}_{O_i=\omega_k}$. If there is only one non-empty set $\textsc{asf}_{O_i=\omega_k}$, that is, if only one potential outcome can be modeled as an actual outcome, the set of complex solution formulas $\textsc{csf}_{\mathbf{O}}$ is identical to $\textsc{asf}_{O_i=\omega_k}$.

To illustrate all four stages, let us now apply CNA to Table 2c. We set $con_t = 0.8$

and $cov_t = 0.9$ and execute the algorithm in the most general manner by not providing an ordering. 2c contains data of type fs, meaning that the values in the data matrix are interpreted as membership scores in fuzzy sets. As is customary for this data type, we use uppercase letters for membership in a set and lowercase letters for non-membership. In the absence of an ordering, the first stage determines the set of potential outcomes to be $\mathbf{O} = \{A, B, C, D, E\}$, that is, the presence of each factor in 2c is treated as a potential outcome. Moreover, all other factors are potential cause factors of every element of $\mathbf{O}$, hence, $\mathbf{C}_A = \{B, C, D, E\}$, $\mathbf{C}_B = \{A, C, D, E\}$, $\mathbf{C}_C = \{A, B, D, E\}$, etc.

To construct the sets of minimally sufficient conditions of the elements of $\mathbf{O}$ in stage 2, CNA first tests the values of single potential cause factors for con_t compliance and then moves on to conjunctions of two, of three, and of four factor values. The resulting sets of minimally sufficient conditions are: $\textsc{msc}_A = \{b{*}C, d{*}E\}$, $\textsc{msc}_B = \{a{*}C, A{*}E, d{*}E\}$, $\textsc{msc}_C = \{A, B, d{*}E\}$, $\textsc{msc}_D = \{E, a{*}C\}$, $\textsc{msc}_E = \{D, A{*}B\}$. Only the elements of $\textsc{msc}_C$ and $\textsc{msc}_E$ can be disjunctively combined to atomic solution formulas that meet cov_t in stage 3: $\textsc{asf}_C = \{A + B \leftrightarrow C\}$ and $\textsc{asf}_E = \{D + A{*}B \leftrightarrow E\}$. For the other three elements of $\mathbf{O}$ the coverage threshold of 0.9 cannot be satisfied. CNA therefore abstains from issuing causal models for A, B and D.

Finally, stage 4 conjunctively combines $\textsc{asf}_C$ and $\textsc{asf}_E$ to the following complex solution formula $\textsc{csf}_\mathbf{O}$, which constitutes CNA's final causal model for Table 2c:

$$(A \ + \ B \leftrightarrow C) * (D \ + \ A{*}B \leftrightarrow E) \qquad con = 0.808;\ cov = 0.925 \tag{7}$$

Two features of this algorithm deserve (re-)emphasis. First, while the computational cores of configurational methods that build models from the top down are constituted by procedures for redundancy elimination turning maximal into minimal sufficient and necessary conditions, all conditions that CNA finds to comply with con_t and cov_t are automatically redundancy-free. That is, CNA directly identifies *minimally* sufficient and necessary conditions, rendering redundancy elimination itself redundant. Second, whereas QCA dichotomizes fs data in a truth table before processing it, CNA processes fs data in the very same vein as cs and mv data, viz. by building all viable conjunctions and disjunctions of potential causes and systematically testing for con_t and cov_t compliance. By directly applying the same algorithm to all configurational data types, CNA renders the detour via truth tables redundant.

Evaluation and Comparison

Before a new method can be applied in real-life studies, it must, on the one hand, be shown that the method correctly analyzes data that, by the method's own standards, faithfully reflect data-generating causal structures, and, on the other, an estimate should be provided of how the method performs under different constellations of data deficiencies.[5] Accordingly, this section reports the results of a series of evaluation tests that follow the template of so-called *inverse searches*, which reverse the order of causal discovery in scientific practice. An inverse search comprises three steps: (1) a causal structure Δ is presupposed, (2) artificial data δ is generated by letting the involved factors behave in accordance with Δ, and (3) δ is processed by a scrutinized method. The method successfully completes the inverse search iff its conclusions are true of Δ.

In what follows, we not only evaluate the performance of the generalized CNA algorithm, but also compare it with QCA's most reliable search strategy, viz. the parsimonious one (Baumgartner and Thiem 2017). To secure the comparability with QCA, the evaluation focuses on CNA's stages 1-3, which search for single-outcome structures (as does QCA) and constitute the method's analytical core.

For the test series, we use the R packages **cna** (Ambühl and Baumgartner 2018), which—in its newest version 2.1.1—implements the generalized CNA algorithm developed here, and **QCApro** (Thiem 2018), which is the most dependable QCA software currently available and additionally offers many valuable tools for method evaluation.[6] The command line interfaces of these R packages facilitate performing and replicating inverse searches—as detailed in the appended replication script. The two packages provide all functions needed for a wide array of trials. The most relevant among these functions are `randomDGS`, which randomly draws data-generating structures Δ from a factor frame $\mathbf{F}$, `allCombs`, which generates the whole space of logically possible configurations of the factors in $\mathbf{F}$, `some` and `sample`, which randomly sample a specified number of cases from a data set, `makeFuzzy`, which fuzzifies the data (e.g. to simulate background noise), `selectCases`, which selects the cases that comply with Δ and randomly adds outlier cases not complying with Δ while ensuring that specified con_t and cov_t thresholds remain satisfied, `submodels`, which generates the set of correct models (as defined in online appendix A), and `cna` and `eQMC`, which analyze the data by means of CNA and QCA, respectively.[7]

Against that background, inverse search trials revolve around the following steps.

1. Use `randomDGS` to draw a data-generating structure Δ from a factor frame $\mathbf{F}$.
2. Use `allCombs` to generate the space α of all logically possible configurations from

[5] For more on CNA's correctness standards see the online appendix A.

[6] As anticipated in footnote 2, the most recent version of the **QCA** R package (Duşa 2007) holds a lot of promise; in particular, because it supplies a new search algorithm, called *CCubes*, that also adopts the bottom-up approach. Moreover, the package provides search parameters (e.g. for solution consistency and coverage) that allow for closely approximating the CNA algorithm developed in this paper. All of that is new and not part of the QCA protocol (yet). Accordingly, none of the conclusions drawn from the ensuing comparison of CNA and QCA have any bearing on CCubes.

[7] For details on the parameters and arguments of these functions as well as their usage, the reader is referred to the reference manuals of **cna** and **QCApro**.

a factor frame $\mathbf{F}' \supseteq \mathbf{F}$.

3. If the data shall be of type fs (e.g. featuring background noise), use `makeFuzzy` to fuzzify α.
4. Use `selectCases` to select, from α, the set of cases δ complying with Δ, and to add outlier cases not complying with Δ as long as con_t and cov_t remain satisfied.
5. If the data shall be fragmentary, use `some` or `sample` to randomly sample a set of cases δ' from δ; otherwise $\delta' = \delta$.
6. If relevant factors shall be omitted from the data, eliminate columns from δ'; otherwise $\delta' = \delta$.
7. Analyze δ' by means of `cna` and `eQMC` at consistency and coverage thresholds of con_t and cov_t.
8. Check whether the outputs of `cna` and `eQMC` feature a correctness-preserving model contained in `submodels`(Δ). The trial counts as passed iff this check is positive.

Depending on the concrete data scenario to be simulated, the particularities and the arrangement of these steps must be suitably varied. More specifically, in order to simulate model *overspecification*, that is, the inclusion of factors in the simulated data that are causally irrelevant in the targeted structure Δ, $\mathbf{F}'$ must be determined to be a proper superset of $\mathbf{F}$ in step 2. Correspondingly, to simulate model *underspecification*, that is, the omission of factors from the data that are causally relevant in Δ, step 6 must be executed. To simulate *data fragmentation* (or limited diversity), the number of cases drawn in step 5 must be smaller than the exhaustive set of cases compatible with Δ. To simulate *inconsistencies* and *imperfect solution coverages*, con_t and cov_t must be set to values below 1 in step 4. The resulting data is of type fs if step 3 is executed, otherwise it is of type cs or mv, depending on what types of factors are chosen for Δ. Finally, to simulate data that are *ideal* by the standards of configurational causal modeling, $\mathbf{F}'$ must be identical to $\mathbf{F}$ in step 2, con_t and cov_t must be set to 1 in step 4, and steps 5 and 6 must not be executed.

We perform a total of 48 different types of tests. In each test type, we randomly draw 30 to 50 data-generating structures (depending on the calculative complexity of the analysis), on which we then perform inverse search trials using both CNA and QCA. 16 of the test types are run on cs data, 16 on fs data, and 16 on mv data. We simulate data scenarios resulting from all logically possible combinations of the following four types of data deficiencies: overspecification (O), underspecification (U), data fragmentation (F) and imperfect solution consistencies and coverages (I). For instance, a scenario as OuFi is one *with* overspecification, *without* underspecification, *with* data fragmentation, and *without* imperfect (i.e. with perfect) consistencies and coverages, OUFI, in contrast, features all four types of deficiencies, while oufi is free of all deficiencies and, hence, results in *ideal* configurational data.

In order to keep the whole test series easily replicable, the complexity of the randomly drawn data-generating structures is kept comparably simple: they feature between three and four exogenous factors and one outcome each. To simulate overspecification, one irrelevant factor is added to the data; and underspecification is simulated by removing one relevant factor. Moreover, in scenarios with data fragmentation, half of the cases that are compatible with the data-generating structure are removed in case of cs or fs data, while in case of *mv* data we remove 80% of the compatible cases; that is, we simulate diversity indices of 0.5 and 0.2, respectively. Finally, in scenarios with imperfect solution consistencies and coverages, the targeted data-generating structures are set to only reach consistencies and coverages of 0.8 in the simulated data.

The bar charts in Figure 1 contrast the correctness ratios obtained in each test type, that is, the ratios of the number of trials passing the test to the total number trials in each test type. For instance, a ratio of 1 means that every trial produced at least one correct model or 0.7 that 70% of the trials did. A number of aspects of our results deserve separate emphasis. First, CNA significantly outperforms QCA in regard to correctness in a number of data scenarios and performs equally well in all others. Second, all data scenarios featuring neither underspecification nor inconsistencies (oufi, Oufi, ouFi, OuFi), which are the scenarios satisfying *configurational homogeneity* (see online appendix A), are faultlessly analyzed by both methods.[8] In *mv* data, even combinations of over- and underspecification do not diminish correctness ratios. This is strong evidence that both CNA and QCA indeed are correct methods of causal inference: if the relevant background assumption concerning data quality, configurational homogeneity, is satisfied, both methods guarantee correct results. Third, as is to be expected, neither method performs without error in the increasingly deficient data scenarios. No method can faultlessly analyze deficient data that do not faithfully reflect data-generating structures. But while QCA's correctness ratios plummet in certain cases, in particular, when over- and underspecification are combined with imperfect consistencies and coverages, CNA maintains reasonable correctness ratios even in those cases.

A proper interpretation of this last finding requires some differentiation. Primarily, it must be emphasized that if CNA has a high and QCA a low correctness ratio in a particular test that does not automatically mean that CNA issues exactly one correct model throughout that test, while QCA keeps misfiring. Rather, it means that CNA does not commit causal fallacies where QCA does. But causal fallacies can be avoided in various ways. For instance, CNA can pass a trial by abstaining from producing any models at all, while QCA issues false models. To assess the frequency of that constellation in our test series, we additionally calculated the ratios of trials within each test type in which CNA and QCA *produce no model at all*—the results are presented in Figure 3 of the online appendix

[8] In regard to the evaluation of QCA, this finding confirms recent results of Baumgartner and Thiem (2017) and contrasts with claims made by Lucas and Szatrowski (2014).

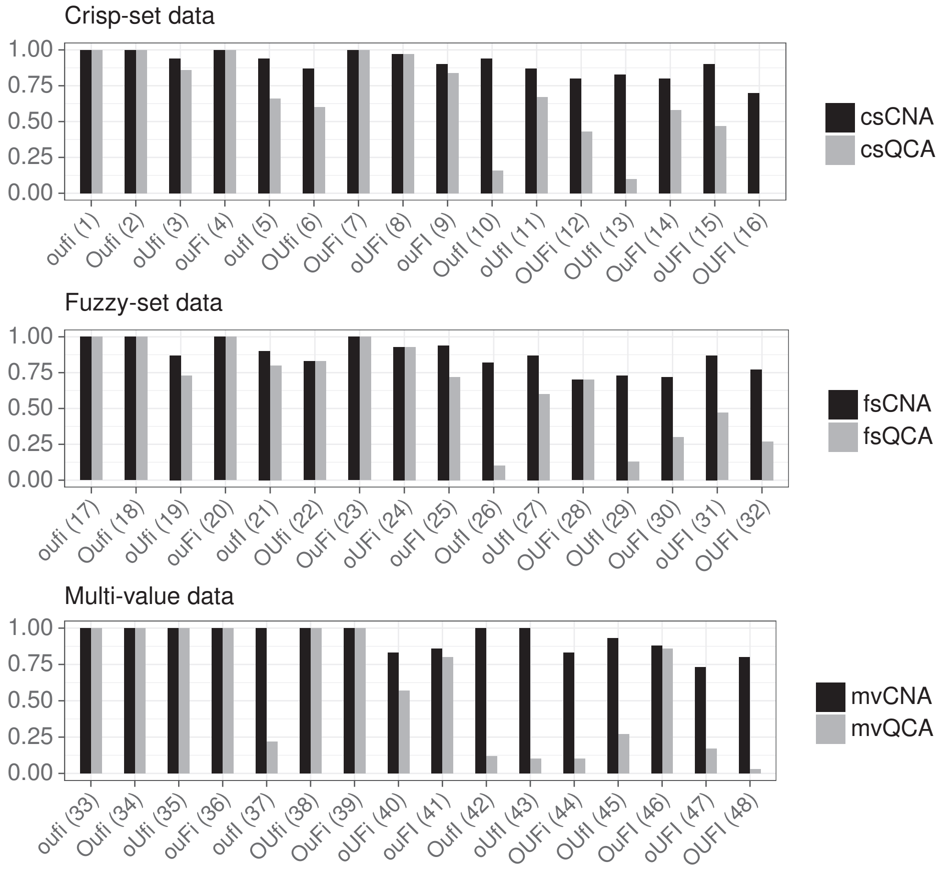

Figure 1. *A comparison of correctness ratios of CNA and QCA for each test type. The latter are listed on the x-axis and numbered in correspondence with the replication script.*

B. It turns out that abstinence from drawing a causal inference is the main reason why CNA outperforms QCA in case of severely deficient *mv* data (i.e. tests 43 to 48) and one reason, among others, in case of deficient fs data (i.e. tests 25 to 32). In those data scenarios, CNA's reliance on both consistency and coverage as authoritative model building criteria prompts CNA to abstain from drawing causal inferences because consistency and coverage thresholds cannot be met. By contrast, QCA, which does not impose coverage thresholds and gives less weight to consistency, continues to draw inferences, committing causal fallacies more often than not.

Alternatively, in cases of data for which multiple equally fitting models exist, a difference in CNA's and QCA's correctness ratios may be due to the fact that CNA more thoroughly uncovers the space of all data-fitting models. Plainly, the more exhaustive the set of alternative models returned by a method, the higher the chances that a correct one is contained therein; and contrapositively, the fewer models a method returns, the lower the chances that one of them is correct. In order to assess the impact of CNA's and QCA's capacities to detect model ambiguities on their overall correctness ratios, Figure 4 (in online appendix B) provides the ratios of trials within each test type in which CNA and QCA *produce more than one model*. For both methods, the ratio of model ambiguities increases with the degree of data deficiency. While QCA has a higher ambiguity ratio than CNA in case of deficient *mv* data (i.e. tests 41, 44, 47, 48), CNA more frequently than QCA issues multiple models in case of deficient *cs* and fs data (i.e. tests 11 to 13, 15, 16, 29, 31, 32). However, when QCA outputs multiple models, often none of them are correct (e.g. in tests 16, 26, 29, 44, 48), whereas when CNA generates multiple models, at least one of them tends to be correct (cf. tests 8, 11 to 16, 23, 28 to 32). Hence, CNA not just issues more models than QCA and, for that reason, has a higher chance of hitting the target on mere quantitative grounds; rather, the quality of its models exceeds the quality of QCA's models.

This is a consequence of CNA's reliance on the bottom-up approach, which more rigorously eliminates redundant factors than QCA's top-down approach. As a result, QCA regularly fails to eliminate irrelevant factors in data scenarios where overspecification is combined with imperfect consistencies and coverages (and possibly other deficiencies).[9] By contrast, the combination of overspecification and imperfect consistencies/coverages does not prevent CNA from reliably eliminating irrelevant factors. Ultimately, this is the main reason why CNA's correctness ratio exceeds QCA's in case of severely deficient *cs* data and a substantial reason in case of deficient fs data.

The question remains how frequently the two methods output a unique model. To answer that, Figure 5 (online appendix B) furnishes the ratios of trials within each test type in which CNA and QCA *produce exactly one model*. This comparison reveals an important difference between QCA and CNA. Exactly one model is QCA's dominant type of output throughout all 48 tests. For CNA, by contrast, this is only the dominant output type in the tests with mild degrees of data deficiency (or none at all). The crucial follow-up question then becomes: what are the ratios of trials such that one unique model is issued that moreover *correctly* reflects the data-generating structure? That question is answered in Figure 6 (online appendix B). Unsurprisingly, QCA's insistence on a unique model has negative effects on the method's overall correctness ratio in all data scenarios with severe deficiencies, for that unique model tends not to be correct in those scenarios. By contrast, no such negative effects result in some data scenarios with only

[9] An interesting exception is test 46, where the combination of overspecification and imperfect consistencies/coverages does not seem to pose a critical problem for QCA.

mild data deficiencies. Notably, in tests 3, 8, 24, 35, and 38, QCA produces a single model more frequently than CNA, but still reaches overall correctness ratios that are comparable to CNA's ratios. Hence, these are tests where QCA draws more precise causal inferences than CNA. This difference in output precision occurs in data scenarios featuring underspecification but no inconsistencies. Those are scenarios where solution coverages tend to be low because relevant factors that are responsible for certain instances of analyzed outcomes are unmeasured, meaning that these instances are not covered by resulting models. As QCA does not impose a coverage threshold, it nonetheless produces outputs, which, since the overall degree of data deficiency is mild, often are correct. CNA, by contrast, imposes authoritative coverage thresholds and is hence disposed to abstain from issuing any models in those cases. By lowering coverage cutoffs, CNA could be induced to behave less cautiously and draw more precise causal inferences in those data scenarios as well. Furthermore, there also exist tests—in particular, tests 10, 17, 18, 37, and 42—in which CNA, even when high coverage standards are enforced, draws more precise inferences than QCA by more frequently issuing one correct unique model and, thus, reaching equal and sometimes considerably better correctness ratios than QCA.

To round off this evaluation, we not only culled correctness, ambiguity, uniqueness, and 'no model' ratios from the 48 test types, but also completeness ratios.[10] The completeness ratio in a test type is the ratio of the number of trials in which a method *completely* uncovers the data-generating structure to the total number of trials. Completeness ratios are presented in Figure 2. As is to be expected, CNA and QCA can only systematically uncover all properties of data-generating structures when the data quality is very high.[11] Moreover, it is clear that in cases of underspecification neither method has a chance of ever finding the complete structure. Apart from these confirmations of theoretical expectations, Figure 2 shows that the completeness ratios of the two methods are very close together across the whole test series, except for the tests 10, 20, 21, 26, 37, and 42 where CNA has a significant edge over QCA.[12] That is, CNA's superior correctness ratios are not offset by overall lower completeness ratios; rather, with regard to completeness, CNA likewise outperforms QCA in a number of data scenarios and performs comparably in all other scenarios.

We end this discussion with some qualifications. The forms of data deficiencies analyzed here do not exhaust the space of possible deficiencies. For instance, all the data we simulated feature evenly distributed case frequencies, that is, different configurations

[10] For details on completeness and its relation to correctness see the online appendix A.

[11] Note, however, that only CNA reliably recovers the complete data-generating structure from ideal data (cf. tests 1, 17, and 33). QCA fails to find the complete structure in 3 out of 50 trials with ideal fs data (test 17). As any method should always be able to infer the complete data-generating structure from ideal data, this is a disturbing finding that calls for further investigation.

[12] There are further tests in which the completeness ratios of the two methods come apart slightly. In tests 5, 23, 27, and 30 CNA's completeness ratio exceeds QCA's, whereas in tests 14, 19, and 25 QCA's ratios exceeds CNA's. We take these differences to be non-significant, as they are subject to variations in the replication seeds.

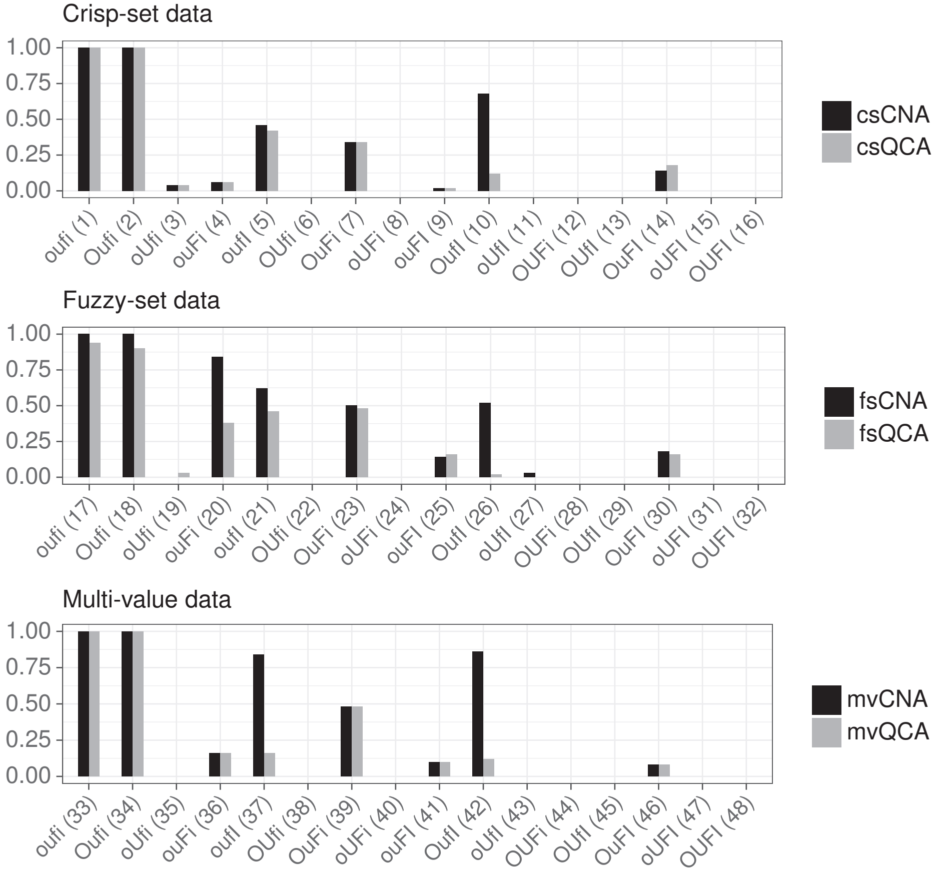

Figure 2. *Completeness ratios for each test type, which are listed on the x-axis and numbered according in correspondence with the replication script.*

are represented by roughly equally many cases. Of course, that is often not the case in real-life data. It is thus an open question how CNA and QCA fare and compare under biased case frequencies. Also, our test series sets consistency and coverage thresholds as well as diversity indices to constant values without exploring how the methods perform under variations of these values. Finally, although we only tested how CNA and QCA perform under various sorts of data deficiencies, we do not intend to suggest that data deficiencies are the only conceivable source of causal fallacies—apart from errors in the internal protocol of a method. There are a host of other sources for causal fallacies: for instance, errors in study designs, faulty background theories, or misapplications of a method. Our test series bracketed all of these problems. All findings reported above must hence be relativized to the particular sources of causal fallacies we chose to simulate.

Conclusion

This paper has generalized Coincidence Analysis (CNA), a configurational comparative method of causal data analysis, for multi-value variables and variables with continuous values from the unit interval that are interpreted as membership scores in fuzzy-sets. Moreover, it has shown in an extended series of benchmark tests that CNA performs both correctly and completely in ideal data scenarios and maintains reliable correctness ratios across a wide range of data deficiencies.

CNA differs from QCA, the currently dominant CCM, in numerous respects. First, CNA not only uncovers single-outcome structures but also structures with multiple outcomes. It is the only CCM custom-built to uncover the Boolean complexity dimension of sequentiality. Second, CNA builds causal models from the bottom up rather than from the top down. Thereby, it renders redundancy elimination (or minimization) itself redundant—which constitutes the algorithmic core of QCA. This reversal of the basic model building approach, on the one hand, allows CNA to abstain from erroneously causally interpreting irrelevant factors in cases of model overspecification and, on the other, permits CNA to directly apply one and the same algorithmic protocol to all data types, without a detour via truth tables. Third, CNA imposes authoritative consistency and coverage cutoffs on causal models (and all their elements), whereas QCA only uses a consistency threshold in truth table generation. In consequence, CNA is much more risk-averse than QCA when it comes to drawing causal inferences, which, in turn, yields that CNA maintains reasonably high correctness ratios even in scenarios featuring severe data deficiencies that cause QCA's ratios to plummet. At the same time, we have seen that this inferential caution does not entail that CNA would fail to completely uncover data-generating structures where QCA succeeds in doing so.

Overall, the generalized version of CNA not only reliably uncovers all Boolean dimensions of causal structures from crisp-set, multi-value, and fuzzy-set data, but also has effective inbuilt controls that abandon an analysis that is too risky due to data deficiencies. In that light, CNA constitutes a powerful methodological alternative for researchers interested in the Boolean dimensions of causality.

References

Ambühl, Mathias, and Michael Baumgartner. 2018. *cna: Causal Modeling With Coincidence Analysis [computer program]. R Package Version 2.1.1.* URL: https://cran.r-project.org/package=cna.

Baumgartner, Michael. 2009. “Inferring Causal Complexity.” *Sociological Methods & Research* 38:71–101.

———. 2013. “A Regularity Theoretic Approach to Actual Causation.” *Erkenntnis* 78:85–109.

———. 2015. “Parsimony and Causality.” *Quality & Quantity* 49:839–856.

Baumgartner, Michael, and Christoph Falk. 2018. “Boolean Difference-Making: A Modern Regularity Theory of Causation.” *PhilSci Archive* http://philsci-archive.pitt.edu/id/eprint/14876.

Baumgartner, Michael, and Alrik Thiem. 2017. “Often Trusted But Never (Properly) Tested: Evaluating Qualitative Comparative Analysis.” *Sociological Methods & Research* doi: 10.1177/0049124117701487.

Cronqvist, Lasse, and Dirk Berg-Schlosser. 2009. “Multi-Value QCA (mvQCA).” In *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques,* edited by Benoît Rihoux and Charles C. Ragin, 69–86. London: Sage Publications.

Downing, Brian M. 1992. *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe.* N.J.: Princeton University Press.

Duşa, Adrian. 2007. “User Manual for the QCA(GUI) Package in R.” *Journal of Business Research* (URL) 60 (5): 576–586.

Goertz, Gary. 2006. *Social Science Concepts: A User’s Guide.* Princeton: Princeton University Press.

Graßhoff, Gerd, and Michael May. 2001. “Causal Regularities.” In *Current Issues in Causation,* edited by W. Spohn, M. Ledwig, and M. Esfeld, 85–114. Paderborn: Mentis.

Hájek, Peter. 1998. *Metamathematics of Fuzzy Logic.* Dordrecht: Kluwer.

Hume, David. 1999 (1748). *An Enquiry Concerning Human Understanding.* Edited by Tom L. Beauchamp. Oxford: Oxford University Press.

Lucas, Samuel R., and Alisa Szatrowski. 2014. “Qualitative Comparative Analysis in Critical Perspective.” *Sociological Methodology* 44 (1): 1–79.

Mackie, John L. 1974. *The Cement of the Universe. A Study of Causation.* Oxford: Clarendon Press.

Ragin, Charles C. 1987. *The Comparative Method.* Berkeley: University of California Press.

———. 2006. “Set Relations in Social Research: Evaluating Their Consistency and Coverage.” *Political Analysis* 14 (3): 291–310.

———. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond.* Chicago: University of Chicago Press.

———. 2009. “Qualitative Comparative Analysis Using Fuzzy Sets (fsQCA).” In *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques,* edited by B. Rihoux and C. C. Ragin, 87–121. Thousand Oaks: Sage.

Suppes, Patrick. 1970. *A Probabilistic Theory of Causality.* Amsterdam: North Holland.

Thiem, Alrik. 2015. “Using Qualitative Comparative Analysis for Identifying Causal Chains in Configurational Data: A Methodological Commentary on Baumgartner and Epple.” *Sociological Methods & Research* 44 (4): 723–736.

———. 2018. *QCApro: Advanced Functionality for Performing and Evaluating Qualitative Comparative Analysis [computer program]. R Package Version 1.1-2.* URL: http://www.alrik-thiem.net/software/.

The European Political Science Association, 2018 *doi:10.1017/psrm.2018.45*

Online Appendix

Causal Modeling with Multi-Value and Fuzzy-Set Coincidence Analysis

MICHAEL BAUMGARTNER and MATHIAS AMBÜHL*

Appendix A: Background on Homogeneity, Model Ambiguities, Correctness

High consistency and coverage scores increase the reliability of the causal models output by CNA but do not guarantee their correctness. To get a clear understanding of the scope, inferential potential, and limitations of CNA, this appendix spells out what it means for the output of CNA to be correct and under what conditions CNA will certainly produce a correct output.

Very generally put, to say that CNA—or any other method—is a *correct* procedure of causal inference means that the causal conclusions it draws from data δ are true of the δ-generating causal structure Δ. This general characterization calls for two specifications. First, no method can be expected to systematically infer true models from deficient data. Whether data meet required quality standards depends on whether they faithfully reflect the causal structure that generated them. But since this structure is unknown in real-life discovery contexts, data quality cannot be assessed analytically but must be imposed by assumption (Cartwright 1989, 55-90). Even heuristics designed to ensure compliance with these assumptions, such as randomization and experimental control, cannot eliminate the risk of insufficient data quality. Accordingly, all procedures of causal inference come with a set of background assumptions, and are only guaranteed to produce correct results provided these assumptions are satisfied.[1]

While in most methodological traditions, the details of these background assumptions are thoroughly investigated and debated, the CCM literature has largely sidestepped this

[*] Appendix A draws on common work with Alrik Thiem, cf. Baumgartner and Thiem (2017b).

[1] For instance, regression analytic methods impose the *Gauss-Markov* assumptions (Gelman and Hill 2007, 45-47), and Bayesian network methods rely on the *Causal Markov* and *Faithfulness* assumptions (Spirtes, Glymour, and Scheines 2000, 29-31).

important issue so far. We cannot exhaustively fill this gap here (which would require a study in its own right), but still want to provide one background assumption—the configurational *homogeneity* assumption—which is sufficient to ensure the correctness of CCMs by ensuring that the analyzed data are not confounded (cf. Baumgartner 2009).[2] Generally put, configurational data are *confounded* iff unmeasured causes change between observed cases in such a way that variations in the outcomes appear to be due to the measured factors, whereas they are actually due to the changing unmeasured causes. Factors that can induce confounding are unmeasured causes of a scrutinized outcome Y that change the value of Y in a way that is not mediated through the measured factors in an analyzed factor frame $\mathbf{F}$, i.e. causes of Y that are connected to Y on at least one causal path that does not go through the elements of $\mathbf{F}$—so-called *off-*$\mathbf{F}$*-path* causes of Y.[3] Changes in off-$\mathbf{F}$-path causes of Y can bring about changes in Y that are erroneously ascribed to a measured factor that merely happens to co-vary with Y without being causally relevant to Y. Configurational data δ are not confounded if all off-$\mathbf{F}$-path causes of Y remain constant across all cases in δ. Accordingly, an assumption that is sufficient to exclude confounding stipulates that δ are *homogenous* in the following sense:

Configurational Homogeneity (CH): Configurational data δ for an outcome Y over a factor frame $\mathbf{F}$ are homogenous iff every off-$\mathbf{F}$-path cause of Y remains constant in all cases in δ.

Requiring δ to be homogenous in this sense amounts to a strong assumption that may be difficult to justify in observational studies. In fact, whenever the coverage of Boolean causal models is non-perfect, it follows that confounders are operative, meaning that **CH** is violated. A violation of **CH**, however, does not entail that causal inferences are impossible or that incorrect models will automatically be generated, it only follows that the correctness of resulting models is no longer guaranteed. Depending on how much risk a researcher is willing to take in a given discovery context, higher or lower degrees of **CH**-violations (e.g. visible in coverage scores) will induce her to abstain from a causal inference. On a par with background assumptions in other methodological frameworks, the function of **CH** is not to determine when causal inferences are possible but merely to *guarantee* the correctness of resulting models. If data δ are homogenous, it follows that all observed differences in the outcomes must be due to variations of the measured factors, which,

[2] We have to leave it to future research to determine whether the homogeneity assumption is also necessary for that purpose, or whether there exist alternative, possibly weaker assumptions that could likewise guarantee CNA's correctness. Moreover, note that data confounding is, of course, not the only data deficiency that can induce causal fallacies, errors of data collection (e.g. measurement error or selection bias) being another common type of data deficiency. For the purposes of this paper, we bracket errors of data collection by assuming that data have been faultlessly collected. Likewise, we do not consider misapplications of the method as a possible source of causal fallacies.

[3] This terminology is derived from Woodward's (2003, 59-60) notion of an *off-path variable*.

in turn, ensures that CNA cannot commit fallacies by ascribing the difference-making relations it uncovers to causal influences of the measured factors.

The second necessary specification of the rough characterization of the correctness criterion concerns the phenomenon of model ambiguities. There often exist multiple causal models that fit data equally well, to the effect that the data underdetermine their own causal modeling. Model ambiguities are a very common phenomenon in all methodological traditions (Simon 1954; Spirtes, Glymour, and Scheines 2000, 59-72; Eberhardt 2013; Baumgartner and Thiem 2017a).[4] Of course, CNA—on a par with any other method—cannot disambiguate what is empirically underdetermined. Rather, it must draw those and only those causal conclusions for which the data *de facto* contain evidence. In cases of empirical underdetermination it must, therefore, render transparent all data-fitting models (and leave the disambiguation up to the analyst). Multiple models in a CNA output are to be interpreted *disjunctively*, meaning that if, say, three models $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$ are returned, CNA determines that the data-generating structure has the form of $\mathbf{m}_1$ *or* that of $\mathbf{m}_2$ *or* that of $\mathbf{m}_3$. Such a disjunction is true iff at least one disjunct is true. Hence, in order for CNA to pass as a correct method of causal inference the data-generating structure must be truthfully reflected by at least one generated model.[5]

Overall, for CNA—or any other CCM—to be a correct method of causal inference it is required that at least one model inferred from homogenous data truthfully reflects the Boolean causal properties of the data-generating structure. More explicitly:

Configurational Correctness (CC): A configurational comparative method $\mathcal{P}$ is a correct procedure of causal inference iff, whenever $\mathcal{P}$ infers a set of models $\mathbf{M}$ from data δ which comply with **CH**, (at least) one model $\mathbf{m}_i \in \mathbf{M}$ satisfies the following four conditions:

(1) all values of exogenous factors contained in $\mathbf{m}_i$ are causally relevant for the corresponding outcome in the δ-generating structure Δ;

(2) if X_1 and X_2 are contained in two different disjuncts in $\mathbf{m}_i$, then X_1 and X_2 are located on two different causal paths in Δ;

[4] As shown by Baumgartner and Thiem (2017a), model ambiguities are much more frequent in configurational causal modeling than is typically acknowledged. In particular, applications of QCA are affected by a widespread practice of model-underreporting, one main reason being that the dominant QCA computer programs—as **fs/QCA** (Ragin and Davey 2016) or **Tosmana** (Cronqvist 2017)—regularly fail to uncover the whole model space, even for ideal data. While **QCA** (Duşa 2007) can avoid this problem if default parameter settings are appropriately tweaked, the only currently available QCA program that recovers the whole model space by default is **QCApro** (Thiem 2018).

[5] An analogous correctness benchmark is implemented in other methodological traditions. Spirtes, Glymour, and Scheines (2000, 81), for instance, require that a correct method returns a pattern of models (i.e. not an individual model) that represents the faithful indistinguishability class of data-fitting models, where a pattern is a disjunction (or class) of models. Similarly, Kalisch et al. (2012, 7), who require their procedures to only report the equivalence class of models in which the true model must lie.

(3) if X_1 and X_2 are contained in the same conjunct in $\mathbf{m}_i$, then X_1 and X_2 are part of the same complex cause in Δ;

(4) if X_1 and X_2 are two links of a causal chain in $\mathbf{m}_i$, then X_1 and X_2 are two links of a causal chain in Δ.

To a model $\mathbf{m}_i$ that truthfully reflects Δ by complying with conditions **CC**(1) to **CC**(4) we refer as a *correct model*.

We claim that CNA is a correct procedure in the sense defined by **CC** and provide substantive evidence for this in the main part of the paper. Two aspects of this claim deserve separate emphasis. First, that CNA is correct does not entail that it infers causal models from every data input. Data may be insufficient to warrant any causal inference. Whenever CNA abstains from an inference, it cannot commit a causal fallacy. By extension, correctness cannot be violated. Configurational causal modeling imposes very high quality standards on the processed data. If these standards are not met, a reliable CCM must refrain from drawing inferences. As detailed in the main part of the paper, CNA adopts a much more risk-averse approach in dealing with data deficiencies than QCA. While the latter does not impose a coverage threshold at all and often causally interprets minimally sufficient conditions that do not meet the consistency threshold, the former uses both consistency and coverage as *authoritative* model building criteria such that, if they are not met, CNA abstains from a causal inference. It is better not to draw a causal inference than to draw a hazardous one.

Second, that CNA is a correct method does not entail that it always *completely* uncovers the data-generating structure Δ. Real-life data tend to be fragmentary, meaning they do not contain all configurations that are empirically possible, that is, compatible with Δ.[6] Fragmentary data may not contain evidence for certain features of Δ, and no method can compensate for lacking evidence. Correctness merely demands that, if CNA outputs a set **M**, then at least one model $\mathbf{m}_i \in \mathbf{M}$ be such that all causal properties represented by $\mathbf{m}_i$ truthfully reflect *some* causal properties of Δ. At the same time, if CNA is given exhaustive data featuring *all* empirically possible configurations, CNA should completely uncover Δ. That is, *completeness* is imposed as a conditional criterion: if CNA is given exhaustive data in compliance with **CH**, the Boolean causal properties represented by at least one model $\mathbf{m}_i \in \mathbf{M}$ truthfully reflect *all* Boolean causal properties of Δ.[7]

[6] *Data fragmentation*, as we use the term here, is related but not synonymous to *limited diversity*, a concept known from QCA (e.g. Ragin 2008, 147-148). QCA-processed data are said to be limitedly diverse iff they do not contain all *logically possible* configurations of the exogenous factors. CNA, by contrast, allows for the factors that are exogenous with respect to some ultimate outcome to be mutually causally dependent, in which case not all logically possible configurations are also empirically possible. Accordingly, we say that data are fragmentary iff they do not contain all *empirically possible* configurations.

[7] In Baumgartner (2009), an assumption of *empirical exhaustiveness* (Pex) is introduced to ensure that CNA-processed data is non-fragmentary and that Δ could be completely uncovered. We dispense with that assumption here. As a result, CNA will not always completely uncover Δ.

Since data fragmentation is ubiquitous in observational studies, procedures employed in this domain usually will only uncover a proper part of Δ. Still, if the data δ are fragmentary, CNA will uncover all those parts of Δ for which δ contain evidence, no fewer and no more. More specifically, although CNA is not unconditionally complete, it is unconditionally *informative* in the following sense: all and only those Boolean causal properties of Δ for which δ contain evidence are truthfully reflected by at least one model $\mathbf{m}_i \in \mathbf{M}$.

References

Baumgartner, Michael. 2009. "Inferring Causal Complexity." *Sociological Methods & Research* 38:71–101.

Baumgartner, Michael, and Alrik Thiem. 2017a. "Model Ambiguities in Configurational Comparative Research." *Sociological Methods & Research* 46 (4): 954–987.

———. 2017b. "Often Trusted But Never (Properly) Tested: Evaluating Qualitative Comparative Analysis." *Sociological Methods & Research* doi: 10.1177/0049124117701487.

Cartwright, Nancy. 1989. *Nature's Capacities and Their Measurement.* Oxford: Clarendon Press.

Cronqvist, Lasse. 2017. *Tosmana: Tool for Small-N Analysis [computer programme], Version 1.53.* Url: http://www.tosmana.net. Trier: University of Trier.

Duşa, Adrian. 2007. "User Manual for the QCA(GUI) Package in R." *Journal of Business Research* (URL) 60 (5): 576–586.

Eberhardt, Frederick. 2013. "Experimental Indistinguishability of Causal Structures." *Philosophy of Science* 80 (5): 684–696.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Kalisch, M., M. Maechler, D. Colombo, M. H. Maathuis, and P. Buehlmann. 2012. "Causal Inference Using Graphical Models With the R Package *pcalg*." *Journal of Statistical Software* 47 (11): 1–26.

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond.* Chicago: University of Chicago Press.

Ragin, Charles C., and Sean Davey. 2016. *fs/QCA: Fuzzy-set/Qualitative Comparative Analysis, version 3.0 [computer program].* Irvine: University of California.

Simon, Herbert A. 1954. "Spurious Correlation: A Causal Interpretation." *Journal of the American Statistical Association* 49 (267): 467–479.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search.* 2nd ed. Cambridge: MIT Press.

Thiem, Alrik. 2018. *QCApro: Advanced Functionality for Performing and Evaluating Qualitative Comparative Analysis [computer program]. R Package Version 1.1-2.* URL: http://www.alrik-thiem.net/software/.

Woodward, James. 2003. *Making Things Happen. A Theory of Causal Explanation.* New York: Oxford University Press.

Appendix B: Additional Test Scores

Ratios of No Models Being Produced

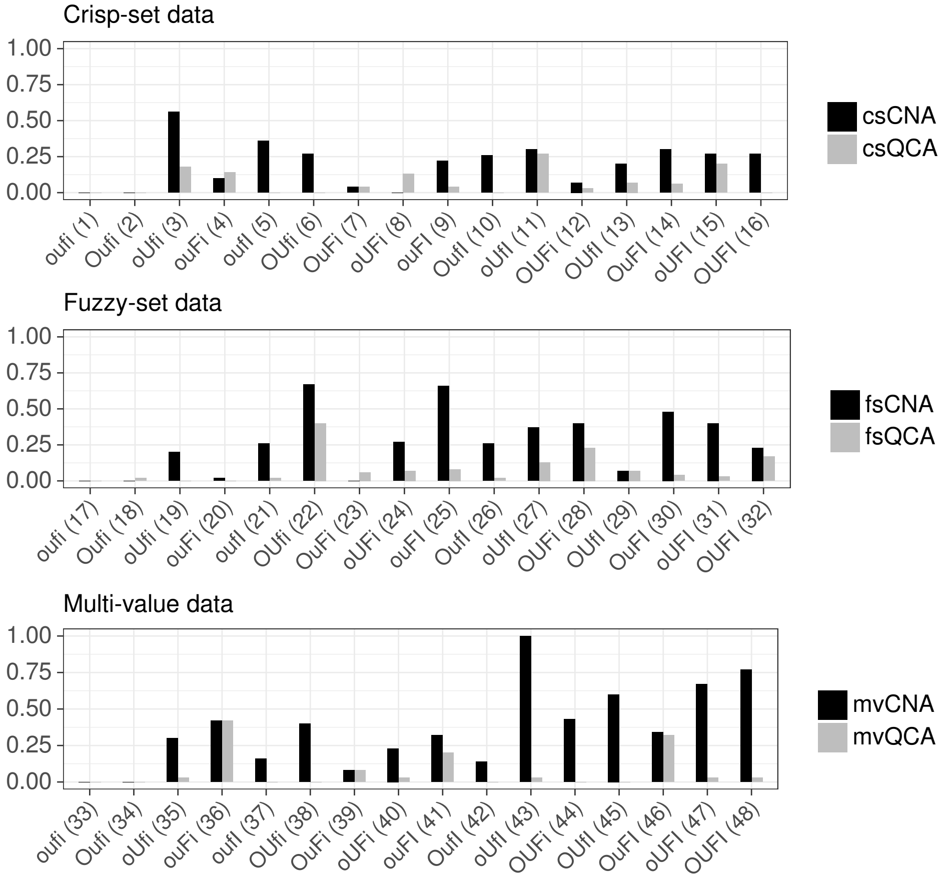

Figure 3. *Ratios of trials in each test type in which no model is produced. The tests are numbered in correspondence with the replication script.*

Ratios of Multiple Models Being Produced

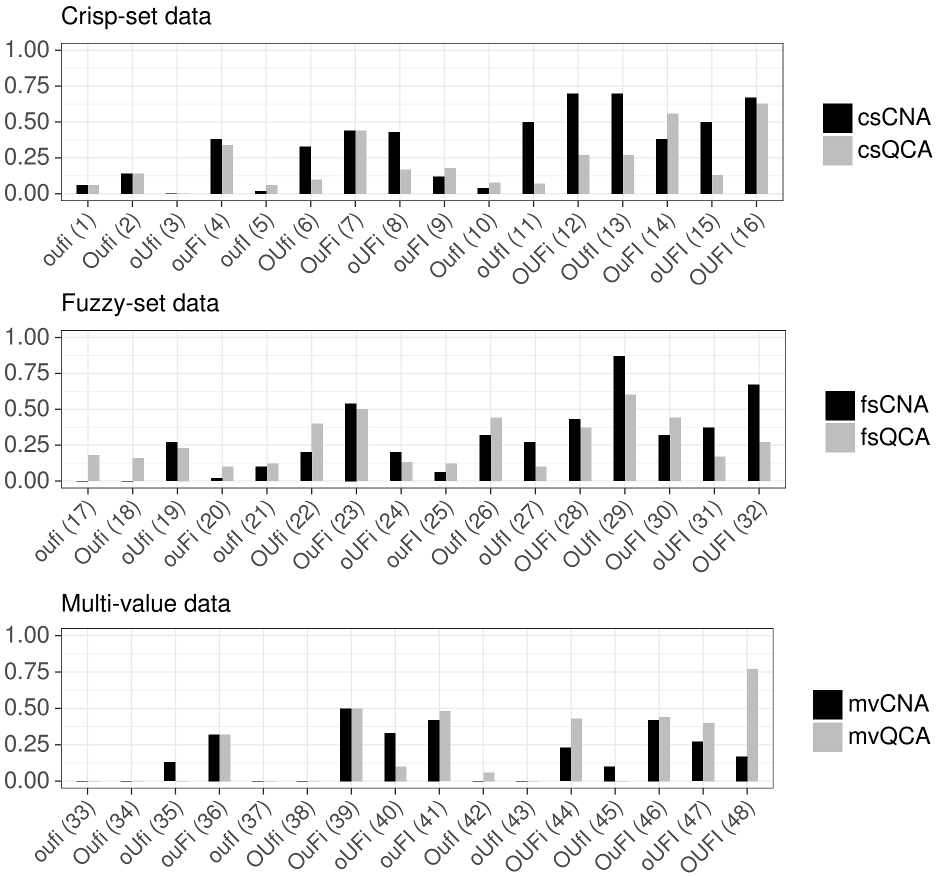

Figure 4. *Ratios of trials in each test type in which more than one model is produced. The tests are numbered in correspondence with the replication script.*

Ratios of Unique Models Being Produced

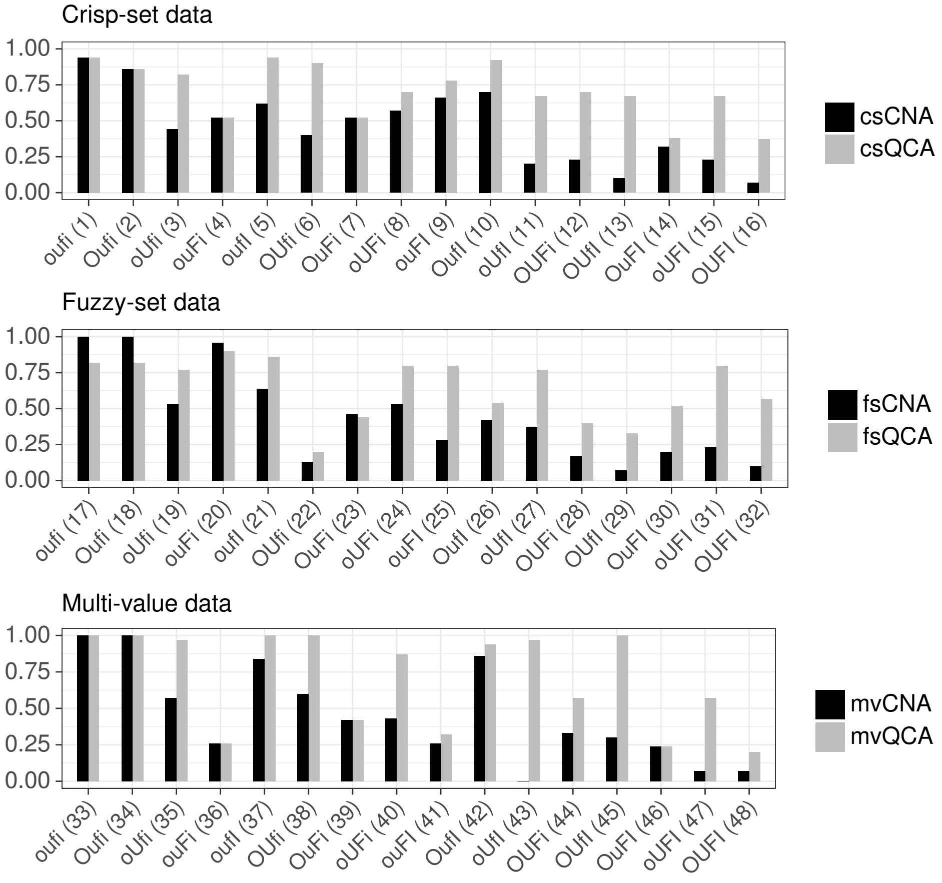

Figure 5. *Ratios of trials in each test type in which one unique model is produced. The tests are numbered in correspondence with the replication script.*

Ratios of Correctness Satisfaction by a Unique Model

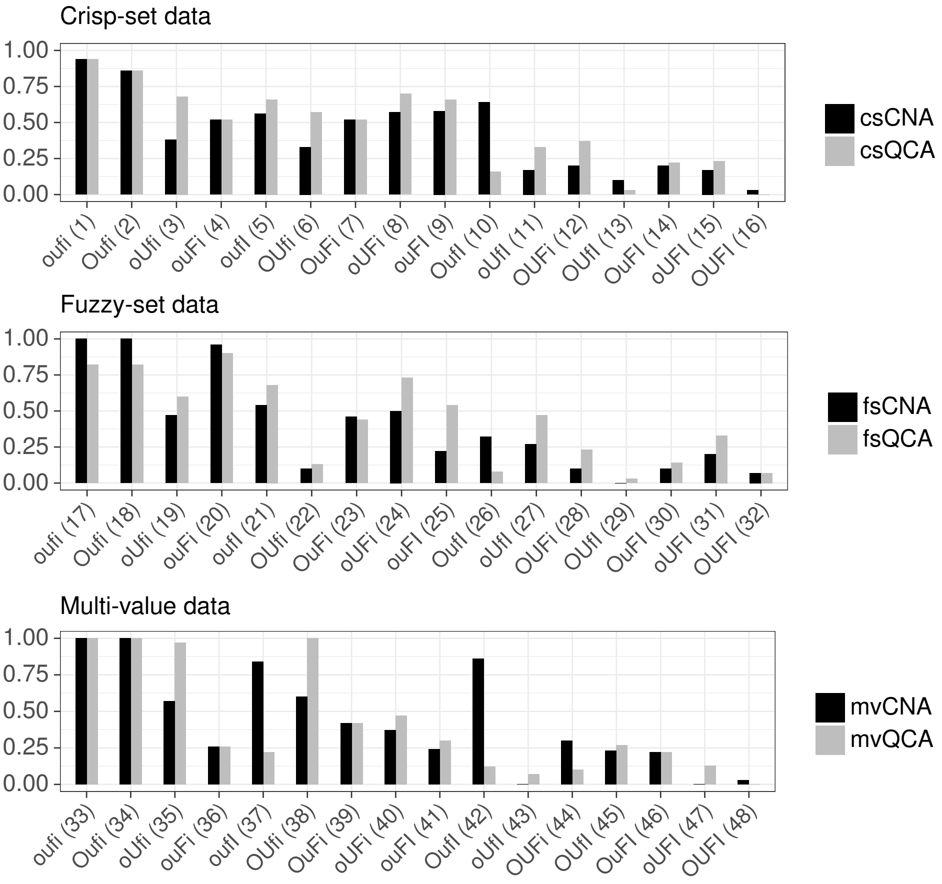

Figure 6. Ratios of trials in each test type in which correctness is satisfied by a unique model, i.e. such that exactly one model is issued which is correct. The tests are numbered in correspondence with the replication script.