Research Article

Luna De Souter*

Evaluating Boolean relationships in Configurational Comparative Methods

https://doi.org/10.1515/jci-2023-0014
received March 14, 2023; accepted August 10, 2023

Abstract: Configurational Comparative Methods (CCMs) aim to learn causal structures from datasets by exploiting Boolean sufficiency and necessity relationships. One important challenge for these methods is that such Boolean relationships are often not satisfied in real-life datasets, as these datasets usually contain noise. Hence, CCMs infer models that only approximately fit the data, introducing a risk of inferring incorrect or incomplete models, especially when data are also fragmented (have limited empirical diversity). To minimize this risk, evaluation measures for sufficiency and necessity should be sensitive to all relevant evidence. This article points out that the standard evaluation measures in CCMs, consistency and coverage, neglect certain evidence for these Boolean relationships. Correspondingly, two new measures, contrapositive consistency and contrapositive coverage, which are equivalent to the binary classification measures specificity and negative predictive value, respectively, are introduced to the CCM context as additions to consistency and coverage. A simulation experiment demonstrates that the introduced contrapositive measures indeed help to identify correct CCM models.

Keywords: Qualitative Comparative Analysis, Coincidence Analysis, INUS-causation, deterministic causation, Boolean causal discovery

MSC 2020: 62D20, 03A10

1 Introduction

Configurational Comparative Methods (CCMs) aim to learn causal structures from datasets. These methods can infer structures with highly complex causal interactions from relatively small datasets without requiring unconditional dependence between individual causes and their effects. Whereas many standard methods of causal structure learning – most notably Bayesian network methods [1] – rely on the faithfulness assumption or relaxed versions of it and assume pairwise dependence between individual causes and their effects [2,3], CCMs do not rely on any version of faithfulness.

They are widely used in various research fields. For instance, the prominent CCM *Qualitative Comparative Analysis* (QCA) has been applied in thousands of studies across disciplines like political science, business research, environmental science, information management, and education science. This includes studies on topics such as investments in renewable energy, UN sanctions, technology transfer, educational poverty, national innovation performance, flood risk management, digital transformation strategies, and youth well-being. The more recently developed CCM *Coincidence Analysis* (CNA) has seen a significant uptick in applica-

* **Corresponding author: Luna De Souter**, Department of Philosophy, University of Bergen, Sydnesplassen 12-13, 5020 Bergen, Norway, e-mail: luna.souter@uib.no

tions in health research, including high impact studies on topics like liver damage, opioid treatment, cancer care, surgical site infection reduction, and obesity treatment.[1]

CCMs aim to derive causally interpretable models from data by exploiting Boolean sufficiency and necessity relationships. One important challenge for these methods is that these Boolean relationships are often not strictly satisfied in real-life datasets, because such datasets usually contain noise. Hence, CCMs infer models that only approximately fit the data, introducing a risk of inferring incorrect or incomplete models, especially when data are also fragmented (have limited empirical diversity). To minimize this risk, evaluation measures for sufficiency and necessity should be sensitive to all relevant evidence: they should reward all evidence in favour of sufficiency and necessity and penalize all counterevidence. Charles Ragin introduced *consistency* and *coverage* to evaluate sufficiency and necessity [5,6]. Roughly speaking, consistency and coverage measure how often sufficiency and necessity are satisfied in a dataset. These measures were initially defined for so-called crisp-set variables, which can only take one of two values for each case in the dataset, e.g. 1 or 0, or true or false, and were then generalized for fuzzy-set variables, which can take any real number between and including 0 and 1 as their values.

Consistency and coverage are currently the standard evaluation measures in CCMs. They have been accepted on common sense grounds without explicit argumentation by Ragin or others. In other methodological frameworks, for instance in binary classification [7] and in association rule learning [8], there has been extensive research on evaluation measures and their properties. While some of the measures studied in those frameworks are potential additions to or alternatives for consistency and coverage, no systematic studies have been conducted on which measures would be best suited for evaluating sufficiency and necessity in CCMs. Moreover, with the exception of Goertz [9] and Schneider and Wagemann [10], which will be discussed in Section 4.4, the limited existing research on improving consistency and coverage in CCMs has focused on issues that only arise when generalizing these measures to the fuzzy-set case.

This article studies measures for evaluating sufficiency and necessity in crisp-set, multi-value, and fuzzy-set CCMs, by drawing inspiration from the field of binary classification. I show an analogy between CCMs and binary classification that enables the application of certain insights from binary classification evaluation to CCM evaluation. This inspires the introduction of two new evaluation measures to CCMs, while also paving the way for transferring more tools from binary classification to CCMs. The two introduced additional measures are sensitive to evidence that is neglected by current consistency and coverage. When evaluating the sufficiency of A for B, consistency only takes into account cases in which A is present. I argue that some cases in which A is absent also constitute relevant evidence for evaluating A's sufficiency for B. Analogously, coverage neglects certain evidence that is relevant for evaluating necessity. A simulation experiment demonstrates that the two new measures indeed help to identify CCM models that are more likely to be correct. The article concludes by providing concrete recommendations for CCM practitioners based on these findings and by outlining approaches for the further improvement of model evaluation and model building in CCMs. R code for implementing these recommendations as well as scripts for reproducing the simulation experiment are available in the online supplementary materials and on GitHub[2].

2 Background

Many disciplines investigate causal structures featuring *conjunctivity* and *disjunctivity*. Conjunctivity means that causes are complex bundles that only bring about the outcome when all components of the bundle are instantiated jointly. It is modelled by Boolean conjunctions (*and*-connections). Disjunctivity means that

[1] For an overview of QCA studies, see the online catalogue https://www.zotero.org/groups/510780/compasss/items/MUZP9FP6/library, and for an overview of CNA studies, see https://www.zotero.org/groups/4567107/coincidence.analysis/items/XXGTHF7F/library. Furthermore, [4] (pp. 18–21) discusses the evolution of the use of QCA between 1994 and 2019.

[2] https://github.com/Luna-De-Souter/Evaluating-Boolean-relationships-in-CCMs.

outcomes can be brought about along alternative causal routes: only one of the causal routes needs to be instantiated to bring about the outcome. Disjunctivity is modelled by Boolean disjunctions (*or*-connections).[3]

Causal inference methods face severe challenges when used for discovering causal structures with conjunctivity and disjunctivity, because causally related variables are often not pairwise dependent in such structures. For instance, when studying causal structures with a phenotypic expression of genes as the outcome, a specific allele at a specific locus can be uncorrelated with the phenotypic expression under investigation, even though, combined with specific alleles on other loci, it is located on a causal path for the phenotypic expression [14,15].[4] As Bayesian network methods and standard regression methods rely on the faithfulness assumption or relaxed versions of it, they struggle to infer such structures – even from ideal data. There do exist protocols for handling interaction effects with two or three exogenous variables for these methods, but such protocols face multicollinearity issues and have tight computational complexity restrictions [17].

Discovering causal structures featuring conjunctivity and disjunctivity requires a method that does not assume faithfulness and that can model complex causal structures of many individual causes by means of complex Boolean *and*- and *or*-functions. However, the space of Boolean functions is vast: for n binary variables, there exist 2^{2^n} functions. So, a method for discovering causal structures with conjunctivity and disjunctivity needs to efficiently search that vast space of possible functions. This is the purpose of CCMs. CCMs infer causal structures as defined by the *INUS*[5] *theory* of causation [18] complemented by additional minimality requirements [19,20]. The INUS theory is a regularity theory of causation that does not require causes and their outcomes to be pairwise dependent and that is ideally suited to account for structures featuring conjunctivity and disjunctivity.

Suppose that CCMs are used on a dataset with crisp-set variables A, B, C, D, and E. I use uppercase italicized letters as shorthand for a variable taking value 1 and lowercase italicized letters for a variable taking value 0. For instance, "A" denotes that A = 1 and "a" denotes that A = 0. Model (1) below is an example of a causal model according to the theory of causation underlying CCMs.

$$A*b + C*D \leftrightarrow E. \tag{1}$$

$A*b + C*D$ is called the antecedent of the model and E is called the outcome of the model. "*" denotes *and* (Boolean conjunction), and "+" denotes *or* (Boolean disjunction). So, the antecedent is a disjunction of conjunctions. For each observation of the variables A, B, C, and D, the antecedent of model (1) is true if A and b are both true or C and D are both true. Otherwise, the antecedent is false. Outcomes in CCM models always consist of a single variable value, in this case E. Model (1) is an example of a single-outcome CCM model. CCMs can also model causal structures with multiple outcomes, each having their own antecedent [21], and can be used on other types of data. While the findings of this article are applicable to multiple-outcome models and can be extended to types of data other than crisp set, this article will, for simplicity, focus on single-outcome CCM models for crisp-set data. Additional explanations for applying the two new evaluation measures to multi-value and fuzzy-set data are provided in Appendix D.

Model (1) says that the antecedent is true *if and only if* (↔) outcome E is true. In other words, the antecedent is *sufficient* and *necessary* for outcome E. The sufficiency of the antecedent for the outcome is denoted by "→" and means that if the antecedent is true, the outcome is also true. The necessity of the antecedent for the outcome is denoted by "←" and means that *only* if the antecedent is true, the outcome is true, i.e. if the outcome is true, the antecedent is also true. To facilitate the comparison with binary classification, I will call a true antecedent *positive* and a false antecedent *negative*, and the same for a true and a false outcome. So, the goal in CCMs is to obtain an antecedent that is positive if and only if the outcome is positive.

According to the theory of causation underlying CCMs [20], model (1) represents a true causal structure only if its antecedent is a minimally necessary disjunction of minimally sufficient conjunctions for E. The

[3] Diverse terms are used to describe conjunctivity and disjunctivity in different disciplines, e.g. "component causation" [11], "conjunctural causation," "equifinality" [12], and "alternative causation" [13].

[4] The vignette of the R package cna [16] presents detailed examples of causal structures that violate faithfulness and provides an in-depth explanation of how CCMs are able to discover the causal structures of those examples.

[5] The acronym *INUS* stands for "*insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition" [18] (p. 62).

antecedent is minimally necessary for E if it is necessary for E and if taking away any of its disjuncts, $A*b$ or $C*D$, would make it lose its necessity for E. For example, if $A*b$ by itself is necessary for E, then $A*b + C*D$ is not *minimally* necessary for E. A conjunction is minimally sufficient for E if it is sufficient for E and if taking away any of its conjuncts would make it lose its sufficiency for E. If a model's antecedent is a minimally necessary disjunction of minimally sufficient conjunctions for its outcome, and if these requirements, in addition to certain further minimality requirements, are satisfied permanently [20], then the variable values in the model's antecedent are causally relevant for the model's outcome. For instance, if the requirements for causal relevance are satisfied by model (1), it implies that A together with b is causally relevant for E, and C together with D is causally relevant for E.

Hence, causal interpretability of CCM models requires that each disjunct of the antecedent is sufficient for the outcome and that the antecedent itself is necessary for the outcome. Moreover, each of these sufficiency and necessity relationships must be minimal. These minimality requirements can, in turn, be expressed in terms of the absence of sufficiency and necessity. For instance, the sufficiency of $A*b$ for E is minimal *if and only if* both $A \rightarrow E$ and $b \rightarrow E$ do *not* hold. Thus, evaluating which sufficiency and necessity relationships do and do not hold is crucial for determining whether a model is causally interpretable.

Sufficiency and necessity are deterministic relationships. Correspondingly, the INUS theory of causation underlying CCMs posits that relationships between causes and their effects are deterministic, meaning that they hold in a dataset if and only if they are satisfied for every case in that dataset. However, these deterministic relationships between causes and their effects do not hold in many real-life datasets due to the presence of *noise*. Noise refers to cases in a dataset that are incompatible with the true causal structure underlying that dataset, i.e. the *data-generating causal structure* (DGCS) [21] (p. 530). For instance, in a dataset with model (1) as its DGCS, a noisy case is a case for which antecedent $A*b + C*D$ is true and outcome E is false or *vice versa*. Noise may result from various data deficiencies, including measurement error and violations of *homogeneity*, meaning that not all confounders remain constant across all cases in the dataset, where "confounder" refers to a cause of the outcome that is located on a path to the outcome containing no measured variables [22] (p. 286).

In addition, deterministic relationships that do not hold in reality may nonetheless hold in some real-life datasets due to the presence of *fragmentation*. Fragmentation means that not all possible configurations of exogenous variables are included in the dataset. So, in a non-fragmented dataset generated from model (1), every combination of values of A, B, C, and D that is compatible with model (1) appears at least once. Fragmentation stems from limitations in data collection. For example, it may be too time consuming to collect enough cases to cover all required combinations of values, or cases featuring specific value combinations may be highly uncommon due to natural limitations in data diversity (e.g. the combination of a rare blood type with a rare medical condition).

For noise-free, non-fragmented datasets, CCMs are *guaranteed* to infer results that are compatible with the true DGCS by simply accepting sufficiency and necessity relationships that are always satisfied in the dataset and rejecting such relationships if they are violated at least once. However, CCMs have been developed to infer reliable causal models from real-life datasets, and such datasets usually feature noise and fragmentation. If there is noise in the data, and thus any case violating a sufficiency or necessity relationship may be a noisy case, then rejecting a sufficiency or necessity relationship when there is at least one case violating it is too strict. Furthermore, data fragmentation entails that there may be cases that are compatible with the DGCS and that violate a considered sufficiency or necessity relationship, but that are not included in the dataset. Therefore, the complete absence of cases violating a considered relationship in a dataset is too weak a requirement for accepting that relationship. That is why CCMs use measures like consistency and coverage for evaluating sufficiency and necessity. Such evaluation measures aim to assess how likely these deterministic relationships are to hold in the noise-free, non-fragmented version of a dataset based on the observed version of that dataset, by rewarding cases that satisfy them and penalizing cases that violate them. Good evaluation measures enable CCMs to obtain reliable results even for datasets featuring certain degrees of noise and fragmentation.

I will propose new evaluation measures for CCMs based on evaluation measures in binary classification. The aim of binary classification is to obtain models that predict the value of a crisp-set (binary) outcome variable. Consider, for example, a binary classification model for outcome E for a dataset with crisp-set variables A, B, C, D, and E. Such a binary classification model should consist of an expression, in whatever form, in terms of variables A, B, C, and D, that is positive for all cases in which E is positive, and that is negative for all cases in which E is negative. To facilitate comparison with CCMs, I call this expression the “antecedent” of a binary classification model. Binary classification differs from CCMs in that the antecedent of a CCM model needs to be a disjunction of conjunctions that satisfies the discussed minimality requirements whereas no general syntactic restrictions are imposed on binary classification models, since these do not strive for causal interpretability. Nevertheless, as for CCMs, the goal in binary classification is to obtain an antecedent that is positive if and only if the outcome is positive.

3 Model evaluation in binary classification and CCMs

3.1 The confusion matrix

In binary classification, the performance of an antecedent on a dataset is often evaluated by means of the confusion matrix, which groups cases in the dataset into four fields based on whether the antecedent and outcome are positive for them [23]. Let X indicate that the antecedent is positive and let x indicate that the antecedent is negative. Analogously, let Y and y stand for a positive outcome and a negative outcome, respectively. Then, a case is grouped in row Y if it has a positive outcome and in row y if it has a negative outcome. It is also grouped in column X if the antecedent is positive for it and in column x if the antecedent is negative for it. The confusion matrix displays, in each of its four fields, the number of cases that are grouped into that field. In CCMs, cases can equivalently be grouped into four fields based on their values for X and Y, where X is a complex antecedent consisting of multiple variable values and outcome Y consists of a single variable value.

In the binary classification confusion matrix shown in Table 1(a), the field with X and Y contains cases for which the antecedent and the outcome are positive. They are correctly predicted to have a positive outcome and are called *True Positives*. TP stands for the number of True Positive cases. In the CCM confusion matrix, as shown in Table 1(b), the field with X and Y likewise contains cases for which the antecedent and outcome are positive. $|X * Y|$ refers to the number of cases with $X * Y$. Cases with X and Y can be called True Positives in the context of CCMs too, because the CCM antecedent correctly is positive for them.

Table 1: Confusion matrix for binary classification (a) and CCMs (b)

(a)

		Antecedent		
		X	x	**ratio**
Outcome	Y	TP	FN	sensitivity
	y	FP	TN	specificity
ratio		PPV	NPV	

(b)

		Antecedent		
		X	x	**ratio**
Outcome	Y	$\|X{*}Y\|$	$\|x{*}Y\|$	coverage
	y	$\|X{*}y\|$	$\|x{*}y\|$	?
ratio		consistency	?	

The field with X and y contains cases for which the antecedent is positive and the outcome is negative. In binary classification, their outcome is erroneously predicted to be positive, and they are called *False Positives*. Cases with X and y can also be called False Positives in CCMs: the antecedent erroneously is positive for such cases, violating its sufficiency for the outcome. Analogously, the field with x and y contains *True Negatives* and the field with x and Y contains *False Negatives*. For False Negatives in CCMs, the antecedent is erroneously negative, violating its necessity for the outcome. So, we can construct confusion matrices for CCM models, just as we can construct them for binary classification models. Cases called True Positive, False Positive, False Negative, and True Negative in binary classification are equivalent to cases, respectively, called $X * Y$, $X * y$, $x * Y$, and $x * y$ in CCMs. For terminological simplicity, those cases are, from now on, described using the CCM terminology in both the binary classification context and the CCM context.

3.2 Evaluation measures

Ratios of cell sizes of the confusion matrix are used as evaluation measures or *parameters of fit*, as they are usually called in CCMs. In binary classification, the ratio of $|X * Y|$ over $|Y|$ is called *sensitivity* and measures the proportion of cases with a positive outcome for which the antecedent also is positive. The ratio of $|x * y|$ over $|y|$ is called *specificity* and measures the proportion of cases with a negative outcome for which the antecedent also is negative. Sensitivity and specificity are considered to be good evaluation measures for binary classification models in many circumstances [23,24]. They are both ratios over the rows of the confusion matrix and are often considered together as a pair, because their combination allows to monitor performance on cases with a positive outcome and cases with a negative outcome separately. This is more informative than relying solely on accuracy, which measures the total proportion of correctly classified cases in the dataset, without separating positive and negative outcome cases. For example, for a dataset in which 90% of cases have positive outcomes, an antecedent that is positive for every case will obtain an accuracy of 0.9. By contrast, considering sensitivity and specificity separately shows that this antecedent performs very poorly on negative outcome cases (specificity equals 0) and is in fact uninformative. So, measuring sensitivity and specificity ensures that antecedents that perform very well on one outcome value while performing very poorly on the other are penalized, even when one outcome value is much more prevalent than the other.

The ratios over the columns, called *positive predictive value* (PPV) and *negative predictive value* (NPV), are also used for evaluating binary classification models. PPV is the ratio of $|X * Y|$ over $|X|$ and measures the proportion of cases for which the antecedent is positive that also have a positive outcome, which amounts to measuring the reliability of positive predictions made by the model. NPV is the ratio of $|x * y|$ over $|x|$ and measures the proportion of cases for which the antecedent is negative that also have a negative outcome, which amounts to measuring the reliability of negative predictions made by the model. Evaluating the reliability of a model's positive and negative predictions is a substantial addition to evaluating the model's performance on positive and negative outcome cases as measured by sensitivity and specificity. For example, a model with 0.9 sensitivity and 0.9 specificity performs reasonably well on both positive outcome cases and negative outcome cases, but if it is applied to a dataset in which only 10% of the cases have a positive outcome, then only 50% of the cases for which its antecedent is positive will actually have a positive outcome. Applied to this dataset, the antecedent's PPV is only 0.5, whereas its NPV is approximately 0.99. This shows the importance of examining PPV and NPV in addition to sensitivity and specificity.

Sensitivity and specificity, on the one hand, and PPV and NPV, on the other, are often used as pairs in binary classification, each pair having its own specific purpose. Yet, in some application domains, PPV and sensitivity are considered together instead.[6] Considering PPV in addition to sensitivity for model evaluation instead of using the more standard pair sensitivity and specificity is motivated by low *prevalence* [24,26–28]. Prevalence is the proportion of cases in the dataset for which the outcome is positive (i.e. $|Y|/N$, where N stands for the total number of cases in the dataset). A model that has relatively many cases with $X * y$ as

[6] In those application domains, PPV and sensitivity are called *precision* and *recall*, respectively [25].

compared to $X * Y$ can still obtain a high specificity when prevalence is low because, as will become clear in Section 4, specificity penalizes cases with $X * y$ in proportion to the very large number of cases with y in such a dataset. So, it is usually more informative to consider PPV, which penalizes cases with $X * y$ in proportion to cases with X, than to consider specificity in low prevalence scenarios.

In CCMs, only one pair of evaluation measures is used: consistency and coverage [6]. They too are ratios of cell sizes of the confusion matrix. Consistency is the ratio of $|X * Y|$ over $|X|$ and is thus equivalent to PPV. Coverage is the ratio of $|X * Y|$ over $|Y|$ and is thus equivalent to sensitivity. Contrary to binary classification, in CCMs, consistency and coverage are used regardless of prevalence. The ratios of $|x * y|$ over $|x|$ (NPV) and of $|x * y|$ over $|y|$ (specificity) are not considered, even though no explicit reasons have been given for neglecting them, and there are currently no terms for these ratios in the context of CCMs. In Section 4, I argue that specificity and NPV *should* be taken into account in CCMs, especially when prevalence is high.

4 Alternative measures for sufficiency and necessity

4.1 Specificity, or contrapositive consistency

Currently, consistency alone is used as a measure for the sufficiency of X for Y ($X \rightarrow Y$) in CCMs. $X \rightarrow Y$ requires that whenever X is positive, Y is also positive. To check whether this requirement is satisfied often enough, consistency ($|X * Y|/|X|$) considers all cases with X and measures the proportion of them for which $X \rightarrow Y$ is satisfied, i.e. for which $X * Y$. So, consistency is an appropriate measure for $X \rightarrow Y$. According to the same reasoning, specificity is an appropriate measure for the sufficiency of y for x ($y \rightarrow x$): $y \rightarrow x$ requires that whenever y is positive (i.e. Y is negative), x is also positive (i.e. X is negative), and specificity ($|x * y|/|y|$) considers all cases with y and measures the proportion of them for which $y \rightarrow x$ is satisfied, i.e. for which $x * y$.

Measuring $y \rightarrow x$ may not seem useful for CCMs, since only $X \rightarrow Y$ and $X \leftarrow Y$ are used in these methods. Yet, $y \rightarrow x$ is logically equivalent to $X \rightarrow Y$. The rule of contraposition allows for rewriting $X \rightarrow Y$ as $y \rightarrow x$ and $y \rightarrow x$ as $X \rightarrow Y$. As consistency measures sufficiency and specificity measures the contrapositive of sufficiency, I call specificity *contrapositive consistency* in the CCM context.

$$\text{Contrapositive consistency} = \frac{|x * y|}{|y|}.$$

To see the logical equivalence of $X \rightarrow Y$ and $y \rightarrow x$, consider their truth table:

X	Y	$X \rightarrow Y$	$y \rightarrow x$
1	1	1	1
1	0	0	0
0	1	1	1
0	0	1	1

$X \rightarrow Y$ and $y \rightarrow x$ are both violated if and only if $X * y$. They differ only in their syntax, not in their content. Carl Hempel famously argued that the equivalence of $X \rightarrow Y$ and $y \rightarrow x$ implies that any case which provides evidence for one of these two formulas also provides evidence for the other [29]. This is called the *equivalence criterion*. If it is violated, then whether a case provides evidence for a claim depends on the syntax of that claim instead of only on content, which is unacceptable according to Hempel. So, specificity, which is the consistency of $y \rightarrow x$, also measures evidence for $X \rightarrow Y$.

As $X \rightarrow Y$ and $y \rightarrow x$ are only false when $X * y$, both consistency and contrapositive consistency are measures that penalize $|X * y|$. Given that $|X * Y| + |X * y| = |X|$, consistency can be rewritten as follows:

$$\text{Consistency} = \frac{|X*Y|}{|X|} = \frac{|X*Y| + |X*y| - |X*y|}{|X|} = \frac{|X| - |X*y|}{|X|} = 1 - \frac{|X*y|}{|X|}.$$

This reformulation shows how consistency penalizes $|X*y|$ in proportion to $|X|$ and, since $|X| = |X*Y| + |X*y|$, consistency decreases when $|X*y|$ increases relative to $|X*Y|$. Contrapositive consistency can be rewritten analogously:

$$\text{Contrapositive consistency} = \frac{|x*y|}{|y|} = \frac{|x*y| + |X*y| - |X*y|}{|y|} = \frac{|y| - |X*y|}{|y|} = 1 - \frac{|X*y|}{|y|}.$$

So, contrapositive consistency penalizes $|X*y|$ in proportion to $|y|$. Since $|y| = |x*y| + |X*y|$, contrapositive consistency decreases when $|X*y|$ increases relative to $|x*y|$.

4.2 Example in favour of contrapositive consistency

An example shows that considering contrapositive consistency in addition to consistency can provide extra information about $X \to Y$. Consider a study on the causes of intrinsic motivation at work, using a dataset consisting of cases corresponding to individual employees of a company Γ. The outcome Y corresponds to *Intrinsic motivation*. Suppose that we want to assess whether $X \to Y$ holds in the noise-free, non-fragmented version of the dataset in order to evaluate the following candidate model:[7]

$$\textit{Competence} * \textit{Challenge} + \textit{competence} * \textit{challenge} \leftrightarrow \textit{Intrinsic motivation}. \tag{2}$$

Table 2(a) shows the confusion matrix for this CCM model for a fictional example dataset. $X \to Y$ has a rather high consistency of $|X*Y|/|X| = 100/(100+10) \approx 0.91$. However, there is not so much evidence for $X \to Y$. In a company such as Γ, where almost every employee is intrinsically motivated, those that have either both competence and challenge or both no competence and no challenge are of course also often intrinsically motivated. Any arbitrary antecedent that is not negatively associated with *Intrinsic motivation* can be expected to reach a high consistency level, of around $|Y|/N = 200/211 \approx 0.95$. Hence, because $|Y|/N$ (prevalence) is high in the dataset corresponding to Table 2(a), a consistency of 0.91 does not amount to strong evidence in favour of $X \to Y$, even though the same consistency would in many datasets be a strong indication that X is a good candidate for being a sufficient antecedent for Y.

Compare this to the confusion matrix in Table 2(b) taken from a fictional study on a different company Θ. The only difference between the two confusion matrices is $|x*y|$: company Θ has many more unmotivated employees that do not have both competence and challenge or both no competence and no challenge. The consistency of $X \to Y$, which is independent of $|x*y|$, is the same for both confusion matrices. But clearly, Table 2(b) gives stronger evidence for $X \to Y$ than Table 2(a). In company Θ, it is no longer the case that almost all employees are intrinsically motivated. Because $|Y|/N = 200/310 \approx 0.65$, it cannot be expected that any arbitrary antecedent that is not negatively associated with *Intrinsic motivation* achieves high consistency. Here, the fact that $X \to Y$ has a consistency of approximately 0.91 is more likely to indicate that $X \to Y$ would hold in the noise-free, non-fragmented version of the dataset. This shows that examining only consistency while neglecting $|x*y|$ does not always tell the whole story about $X \to Y$.

Now, consider contrapositive consistency for both confusion matrices: in Table 2(a), contrapositive consistency $=|x*y|/|y| = 1/11 \approx 0.09$, whereas in Table 2(b), contrapositive consistency $=100/110 \approx 0.91$. So, in the case where high consistency does not amount to strong evidence for $X \to Y$, contrapositive consistency is low, while in the case where high consistency *does* amount to strong evidence for $X \to Y$, contrapositive

[7] While the example datasets presented in this article are fictional, the research question and model in the example are based on [30,31]. Model (2) represents intrinsic motivation according to *flow theory*. This theory says that a state of high motivation, called "flow," occurs when competence and challenge are either both high or both low. This entails that there is no unconditional correlation between competence and intrinsic motivation or between challenge and intrinsic motivation in noise-free, non-fragmented data for this model, i.e. faithfulness is violated. See the vignette of the R package cna [16] for a description of how CCMs are able to infer models like model (2).

Table 2: Confusion matrices for company Γ (a) and company Θ (b)

(a)

		Antecedent	
		X	x
Outcome	Y	100	100
	y	10	1

(b)

		Antecedent	
		X	x
Outcome	Y	100	100
	y	10	100

consistency is high. It follows that taking into account contrapositive consistency in addition to consistency can be very useful for judging how much evidence there is for $X \rightarrow Y$.

The example demonstrates that the strength of evidence for $X \rightarrow Y$ given by a certain consistency level can vary depending on prevalence. For low prevalence, high consistency is informative for judging that $X \rightarrow Y$ would hold in the noise-free, non-fragmented version of the dataset, but for high prevalence, consistency is not very informative. This aligns with the different uses of the pair PPV (consistency) and sensitivity (coverage), on the one hand, and the pair sensitivity and specificity (contrapositive consistency), on the other, in binary classification. When prevalence is low, it is interesting to consider PPV and sensitivity, but when prevalence is high, other measures should be taken into account too.

4.3 NPV, or contrapositive coverage

For measuring the necessity of X for Y ($X \leftarrow Y$), NPV or, as I will call it, *contrapositive coverage* should be considered in addition to coverage. Coverage is an appropriate measure for $X \leftarrow Y$ because it considers all cases with Y and measures the proportion of them for which $X \leftarrow Y$ is satisfied, i.e. for which $X * Y$. Analogously, contrapositive coverage is an appropriate measure for $y \leftarrow x$, because it considers all cases with x and measures the proportion of them for which $y \leftarrow x$ is satisfied, i.e. for which $x * y$. $y \leftarrow x$ is logically equivalent to $X \leftarrow Y$: both are violated if and only if $x * Y$. So, coverage and contrapositive coverage measure logically equivalent expressions. Because $X \leftarrow Y$ and $y \leftarrow x$ are false when $x * Y$, both coverage and contrapositive coverage penalize $|x * Y|$. Coverage can be reformulated as follows:

$$\text{Coverage} = \frac{|X * Y|}{|Y|} = \frac{|X * Y| + |x * Y| - |x * Y|}{|Y|} = \frac{|Y| - |x * Y|}{|Y|} = 1 - \frac{|x * Y|}{|Y|}.$$

This makes clear that coverage penalizes $|x * Y|$ in proportion to $|Y|$. Analogously, contrapositive coverage can be reformulated to show that it penalizes $|x * Y|$ in proportion to $|x|$:

$$\text{Contrapositive coverage} = \frac{|x * y|}{|x|} = \frac{|x * y| + |x * Y| - |x * Y|}{|x|} = \frac{|x| - |x * Y|}{|x|} = 1 - \frac{|x * Y|}{|x|}.$$

An example analogous to the one given for contrapositive consistency in Section 4.2 demonstrates that cases with $x * y$ can provide relevant additional information for evaluating $X \leftarrow Y$. Consider the confusion matrix shown in Table 3(a) for model (2) and a fictional example dataset representing employees in a company Λ. X again corresponds to the antecedent and Y to the outcome of model (2). We want to evaluate whether $X \leftarrow Y$ would hold in the noise-free, non-fragmented version of the dataset. In Table 3(a), $X \leftarrow Y$ has a rather

Table 3: Confusion matrices for company Λ (a) and company Φ (b)

(a)

		Antecedent	
		X	x
Outcome	Y	100	10
	y	100	1

(b)

		Antecedent	
		X	x
Outcome	Y	100	10
	y	100	100

high coverage of $|X * Y|/|Y| = 100/(100 + 10) \approx 0.91$. But there is not much evidence in the data for $X \leftarrow Y$. As almost all employees of Λ have either both competence and challenge or both no competence and no challenge, X can be expected to reach high coverage, of around $|X|/N \approx 0.95$, for any arbitrary variable value in the data. Thus, because $|X|/N$ is high, a coverage of 0.91 does not amount to strong evidence in favour of $X \leftarrow Y$.

Compare this to the confusion matrix for yet another company Φ given in Table 3(b). The only difference between the two confusion matrices is $|x * y|$: company Φ has more unmotivated employees without the antecedent of model (2). Hence, the coverage of $X \leftarrow Y$ is the same for both tables. But clearly, compared to Table 3(a), Table 3(b) gives stronger evidence for $X \leftarrow Y$. The additional employees with $x * y$ in company Φ make it implausible that model (2)'s antecedent has high coverage for any arbitrary variable value in the dataset. It follows that for company Φ's dataset, $X \leftarrow Y$'s coverage of approximately 0.91 is more likely to indicate that $X \leftarrow Y$ is satisfied in the noise-free, non-fragmented version of the dataset.

Now, consider contrapositive coverage for both confusion matrices: in Table 3(a), contrapositive coverage $= |x * y|/|x| \approx 0.09$, whereas in Table 3(b), contrapositive coverage ≈ 0.91. So, in the case where high coverage does not amount to strong evidence for $X \leftarrow Y$, contrapositive coverage is low, while in the case where high coverage *does* amount to strong evidence for $X \leftarrow Y$, contrapositive coverage is also high. Particularly when $|X|/N$ is high, we can obtain extra information about the strength of evidence for $X \leftarrow Y$ by considering contrapositive coverage in addition to coverage.

This is analogous to contrapositive consistency, which, as discussed in Section 4.2, is particularly informative for evaluating $X \rightarrow Y$ when $|Y|/N$ is high. However, whereas $|Y|$ depends only on the dataset, $|X|$ also depends on the model. Therefore, as model-independent evaluation standards are preferable for comparing different models for the same dataset, it is less straightforward to assess whether contrapositive coverage is informative than to assess whether contrapositive consistency is informative. Fortunately, it is also possible to judge when contrapositive coverage is more informative than coverage based on $|Y|/N$: if consistency and coverage are reasonably high, e.g. at least around 0.7, then $|X|/N$ is close to $|Y|/N$, as shown in Appendix A. Models that are built using CCMs have reasonably high minimum consistency and coverage, specified by the researcher. So, for evaluating models built by CCMs, we may assume that $|X|/N$ is close to $|Y|/N$. It should thus suffice to attach extra importance to contrapositive coverage when $|Y|/N$ is high and to attach extra importance to coverage when $|Y|/N$ is low.

4.4 The relevance of $x * y$ in CCMs

Before presenting a simulation experiment which demonstrates the usefulness of contrapositive consistency and contrapositive coverage for evaluating CCM models, I show that the existing CCM literature does not contain any objections against these contrapositive measures and that no evaluation measures with the same rationale or specification as contrapositive consistency or contrapositive coverage have been proposed for CCMs before.

Contrapositive consistency and contrapositive coverage depend on $|x * y|$, so the examples in favour of contrapositive consistency and contrapositive coverage provided in Sections 4.2 and 4.3 demonstrate that evaluation measures that take into account $|x * y|$ can be informative for evaluating $X \rightarrow Y$ and $X \leftarrow Y$. However, Charles Ragin's consistency measure, which is presented as a self-contained sufficiency measure [6], is completely independent of $|x * y|$, implying that $|x * y|$ is uninformative for evaluating $X \rightarrow Y$. The same applies to coverage. Ragin claims that a case such as $X * Y$ "is clearly more relevant to the set-theoretic argument than a consistent case with low scores" (p. 295), where "a set-theoretic argument" refers to expressions such as $X \rightarrow Y$ and where "a consistent case with low scores" refers to cases such as $x * y$. In the same article, Ragin claims that considering cases with $X * Y$ as relevant and cases with $x * y$ as irrelevant for evaluating expressions such as $X \rightarrow Y$ aligns with "commonsense thinking" (p. 295), but he does not give further justification for this.

Seawright [32] started a discussion on the relevance of cases with $x * y$ by arguing that not only cases with Y but also some cases with y are relevant for evaluating $X \leftarrow Y$. He proposes a method for evaluating $X \leftarrow Y$ based on this idea. Clarke [33] and Braumoeller and Goertz [34] argue against Seawright's proposal using technical arguments, which however do not refute the relevance of cases with $x * y$ itself.[8] Goertz [9] and Schneider and Wagemann [10] have further investigated the relevance of cases with $|x * y|$ and have proposed measures that do depend on these cases. However, as I argue in the remainder of this section, these proposals fall short of appropriately addressing the issues caused by the neglect of $|x * y|$ pointed out in this article. I follow Goertz and Schneider and Wagemann in focusing the discussion on $X \leftarrow Y$. Furthermore, to facilitate comparison with crisp-set contrapositive coverage, I will only consider crisp-set data, even though Goertz's and Schneider and Wagemann's measures are most often applied to fuzzy-set data. The fuzzy-set formulations of these measures as well as derivations of their crisp-set variants, which differ substantially from contrapositive coverage, can be consulted in Appendices B and C.

Goertz [9] introduces his measure with the aim of evaluating the importance of necessary conditions. According to Goertz, this importance can be analyzed partly in terms of trivialness: if there are no or almost no cases with x in the dataset, then X is always or almost always positive, making it a trivial necessary condition. Goertz claims that trivial necessary conditions have little or no importance (p. 90), and he proposes a measure for the *nontrivialness* of necessary conditions. This measure depends on $|x * y|$: roughly speaking, necessary conditions with higher $|x * y|$ are more non-trivial. However, the rationale for Goertz's measure is different from that of contrapositive coverage. While contrapositive coverage is an additional evaluation measure for $X \leftarrow Y$, Goertz's nontrivialness measure evaluates the importance of X under the presupposition that $X \leftarrow Y$ is satisfied (pp. 90–91).

Schneider and Wagemann connect trivialness to the size of $|X|$. They build on Goertz's discussion of trivialness and add a substantial motivation for improving the evaluation of antecedents with high $|X|$ as necessary antecedents: if $|X|$ is high, X "automatically passes the formal requirement of being classified as a necessary condition for whatever the outcome set is. It does so not by virtue of its substantive or causal relevance but by its empirical distribution" [10] (pp. 233–234). In other words, if $|X|$ is high, X does not obtain the required coverage by virtue of the truth of $X \leftarrow Y$ in the noise-free, non-fragmented version of the dataset but by virtue of the size of $|X|$. To deal with this issue, Schneider and Wagemann propose their own measure, *Relevance of Necessity*, as an addition to coverage.

[8] Seawright [32] proposed a method for evaluating $X \leftarrow Y$ that gives as much weight to cases with $x * y$ as to cases with $X * Y$ based on his argument for the relevance of some cases with y. Clarke [33] (p. 195) argued against Seawright by showing that cases with $X * Y$ and $x * y$ cannot in general be equally relevant for $X \leftarrow Y$. Even though this constitutes a justified critique of Seawright's method, it does not imply that cases with $x * y$ are in general irrelevant, nor that they are less relevant than cases with $X * Y$. Braumoeller and Goertz [34] also provide arguments against specific aspects of Seawright's method, but they explicitly agree with Seawright that cases with $x * y$ *are* relevant for evaluating claims like $X \rightarrow Y$ and $X \leftarrow Y$.

The motivation for Relevance of Necessity is to some extent similar to my motivation for introducing contrapositive coverage, but there is an important distinction: while Schneider and Wagemann state that *every* X with high $|X|$ obtains the required coverage by virtue of the size of $|X|$, I merely claim that it is *possible* for an X with high $|X|$ to obtain the required coverage by virtue of the size of $|X|$. This distinction leads to a crucial difference between Relevance of Necessity and contrapositive coverage: while Relevance of Necessity is low for any antecedent with $|X|$ substantially higher than required for satisfying $X \leftarrow Y$, contrapositive coverage is still high for such antecedents if a high proportion of cases with x satisfies $X \leftarrow Y$. By penalizing $|x * Y|$ in proportion to $|x|$, contrapositive coverage assesses whether an antecedent obtained the required coverage by virtue of $X \leftarrow Y$ being likely to hold in the noise-free, non-fragmented version of the dataset instead of by virtue of the size of $|X|$.

A few other CCM researchers have formulated critiques of consistency and coverage and, correspondingly, have proposed alternative variants of these measures, e.g. [35–38]. Ragin himself introduces the so-called PRI measure as a “more refined and conservative measure of consistency” [39] (p. 50). These proposals concern problems that arise exclusively for fuzzy-set data, not for crisp-set or multi-value data. My proposal in the current article applies not only to fuzzy-set data but also to crisp-set and multi-value data and points to the fact that consistency and coverage underestimate the evidence in favour of $X \rightarrow Y$ and $X \leftarrow Y$ given by cases with $x * y$. By contrast, in [35–39], all positive influence of cases such as $x * y$ on the proposed measures is *avoided*.[9]

In sum, I have argued that contrapositive consistency should be considered in addition to consistency for evaluating sufficiency and that contrapositive coverage should be taken into account in addition to coverage for evaluating necessity in CCMs. Examining contrapositive consistency and contrapositive coverage is especially important when $|Y|/N$ is high, because in that case, high consistency and coverage are less informative for evaluating $X \rightarrow Y$ and $X \leftarrow Y$. In the next section, I demonstrate with a simulation experiment that when prevalence is high, considering contrapositive consistency and contrapositive coverage in addition to consistency and coverage indeed helps to identify models that are more likely to be correct.

5 Simulation experiment

5.1 Set-up

The purpose of the simulation experiment is to compare the correctness of CCM models that meet certain contrapositive consistency and contrapositive coverage thresholds to the correctness of models that do not meet these thresholds, for high-prevalence datasets. A CCM model is correct if and only if it does not make any false causal relevance ascriptions, meaning that it is a submodel of the DGCS underlying the dataset being analyzed [40] (pp. 8–9). For instance, the model $A{*}D + C{\leftrightarrow}E$ is incorrect with respect to DGCS $A{*}B + C{*}D{\leftrightarrow}E$ due to a false causal relevance ascription to D in conjunction with A. The correctness of a group of CCM models is measured by the proportion of correct models in that group. Rather than demonstrating the optimal way in which contrapositive measures can be used in CCM model evaluation, this experiment provides additional support for adopting these contrapositive measures, by illustrating one way in which their use can benefit the quality of CCM models.

The experiment consists of a series of 50.000 inverse search trials in which noisy and fragmented datasets are generated from randomly drawn DGCSs, after which a CCM analysis is performed on each of these datasets with the aim of recovering the corresponding DGCS. The formation of the random DGCS in each trial begins by randomly generating a disjunction of conjunctions comprising up to six unique variables and with two to four disjuncts each consisting of two to four conjuncts. This expression is then minimized to a well-formed

[9] For fuzzy-set variables and when evaluating $X \rightarrow Y$, “cases such as $x * y$” are cases where the antecedent and outcome have values close to 0 and where the value of the antecedent is lower than or equal to the value of the outcome.

CCM antecedent, which is paired with an outcome to create the DGCS. Subsequently, a dataset composed of all cases compatible with the DGCS is generated, after which a portion of these cases is removed to simulate fragmentation. The proportion of cases to be removed is sampled uniformly from the interval 0.2–0.5. From the resulting fragmented and noise-free dataset, 100 observations are sampled with replacement, and then a portion of these cases are replaced by cases incompatible with the DGCS to simulate noise. The proportion of cases to be replaced by noisy cases is sampled uniformly from the interval 0.05–0.3. Finally, datasets with prevalence varying from 0.6 to 0.9 in steps of 0.05 are created for each of the 50.000 DGCSs by appropriately duplicating cases with positive or negative outcomes while preserving the proportions of fragmentation and noise. This allows for conducting inverse search trials at systematically varied prevalence. In line with the conclusions presented in Section 4, the experiment only includes prevalence levels of 0.6 and above. Additional research is needed to explore the effectiveness of the contrapositive thresholds on datasets with lower prevalence.

The inverse search trials are conducted using the CCM CNA [16]. To ensure generalizability to CNA studies with various settings, the trials are conducted at systematically varied consistency and coverage thresholds. The consistency threshold specifies the minimum consistency required for accepting a conjunction as sufficient, and the coverage threshold specifies the minimum coverage required for accepting a disjunction of conjunctions as necessary. The consistency and coverage thresholds are varied independently between 0.7 and 0.85 in steps of 0.05, giving rise to 16 different consistency-coverage combinations.

CNA returns all CCM models that meet the specified consistency and coverage thresholds. Its output is empty if no models meet these thresholds. After conducting the CNA analyses, the experiment proceeds by introducing contrapositive consistency and contrapositive coverage thresholds. The contrapositive consistency threshold is set equal to the consistency threshold, and the contrapositive coverage threshold is set equal to the coverage threshold for the given trial. The returned models that meet both contrapositive thresholds, *high contrapositive models*, are separated from those that do not, *low contrapositive models*, and the correctness of each resulting group of models is recorded. If one or both groups are empty in a trial, then no correctness is recorded for that group in that trial. For each prevalence-consistency-coverage combination, the correctness per group is averaged across the trials, leading to a comparison of the average correctness of high contrapositive models to that of low contrapositive models *per output*. This avoids a disproportionate influence of outputs containing many models when testing whether selecting high contrapositive models helps to obtain correct causal models.

To usefully interpret the results of the simulation experiment, it is important to consider not only the correctness of models, but also their complexity and degree of completeness. Correctness refers to the absence of false causal relevance ascriptions, and can thus be achieved more easily by making fewer causal relevance ascriptions, but this can also negatively impact the overall quality of the model by decreasing its ability to fully reflect the causal structure underlying the data. To account for this trade-off, the complexity of the models is recorded as well. Complexity is defined as the number of variable value appearances in the antecedent of the model. For example, model $A*B + A*c \leftrightarrow E$ has complexity 4 due to the two appearances of A, and one appearance each of B and c. In addition, the degree of completeness is calculated as the model's complexity in proportion to the complexity of the DGCS. This allows to measure the extent to which a model gives a complete picture of the underlying causal structure, which is a better reflection of model quality than complexity alone. The average complexity and average degree of completeness are recorded per group of models across all inverse search trials. If a group has no correct models in a trial, no complexity or completeness results are recorded for that group in that trial.

5.2 Results

Figure 1 shows the average correctness per output of low contrapositive models (black bars) and high contrapositive models (grey bars). In some consistency-coverage-prevalence settings, only very few or no trials resulted in an output containing at least one high contrapositive model. Missing bars in Figures 1–3 indicate

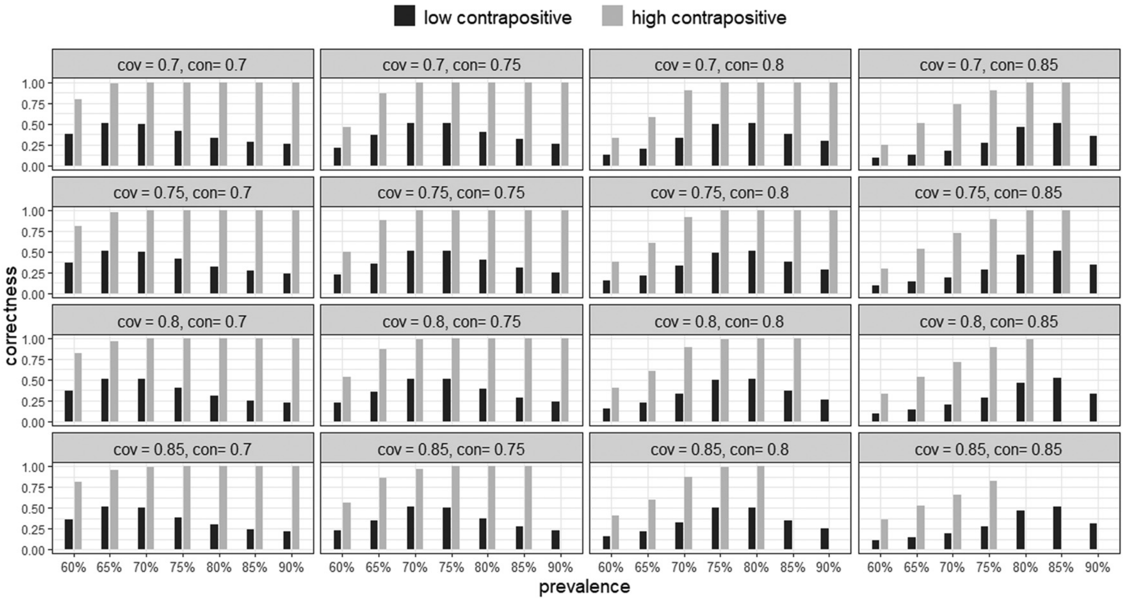

Figure 1: Correctness means (missing bars indicate that mean is based on <20 trials).

settings and groups for which the mean is calculated over fewer than 20 trials; they do *not* represent means that are equal to 0. Standard errors of the means in Figure 1 are between 0 and 0.032 for high contrapositive models and between 0.0006 and 0.0015 for low contrapositive models. Varying prevalence levels are presented on the *x*-axes within the plots, and consistency thresholds and coverage thresholds are presented in different columns and rows, respectively.

Clearly, high contrapositive models have a much higher correctness than low contrapositive models. When consistency is not much lower than prevalence, virtually all high contrapositive models are correct. Correctness of low contrapositive models reaches a peak at prevalence levels close to the consistency threshold and goes down again as prevalence increases above consistency. Strikingly, the same pattern is not found for high contrapositive models: whereas correctness goes up for these models as prevalence increases towards the

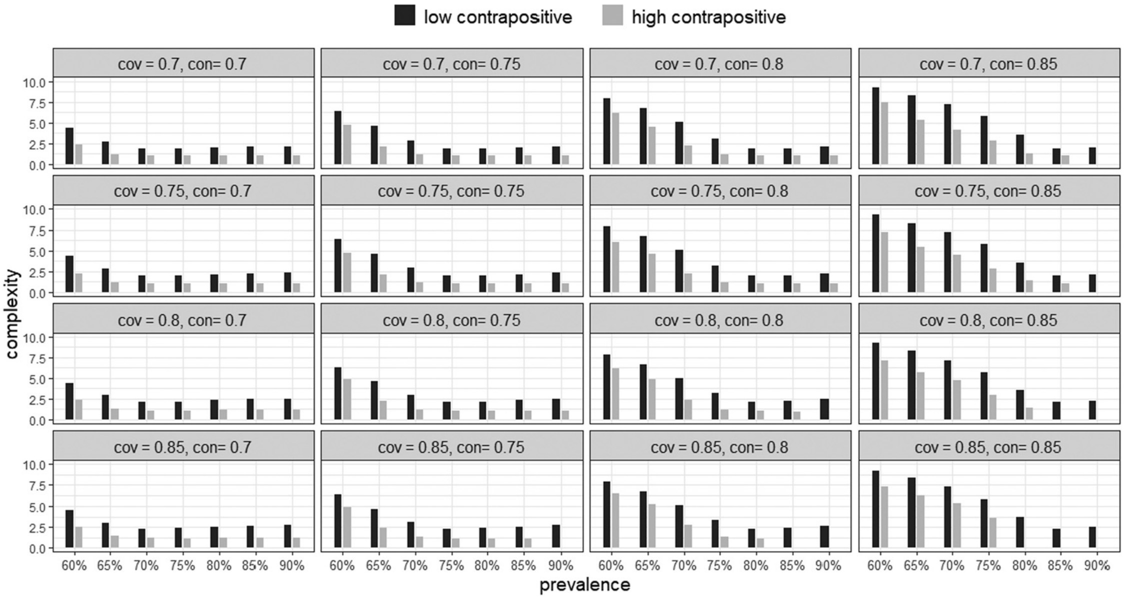

Figure 2: Complexity means (missing bars indicate that mean is based on <20 trials).

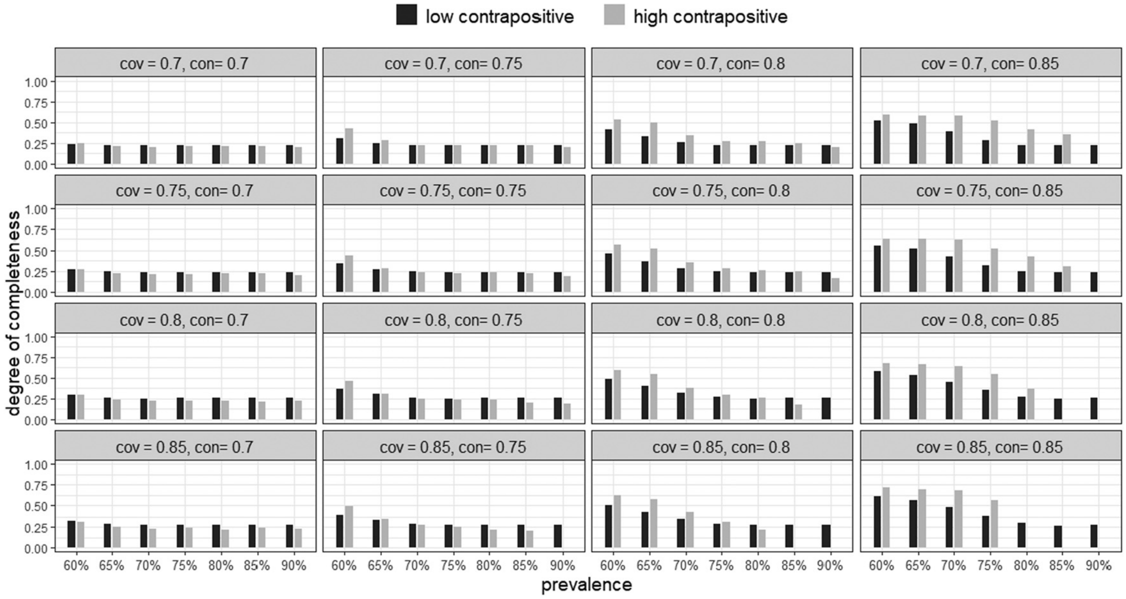

Figure 3: Completeness means (missing bars indicate that mean is based on <20 trials).

consistency threshold, it does not go down again as prevalence increases above consistency. Note also that, while patterns in correctness and prevalence depend substantially on the consistency threshold, they are virtually independent of the coverage threshold.

Figure 2 shows that high contrapositive models have a lower complexity than low contrapositive models. By contrast, Figure 3 illustrates that high contrapositive models do not have a lower degree of completeness than low contrapositive models and that, at high consistency thresholds, the degree of completeness of high contrapositive models is even higher than that of low contrapositive models. This contrast is surprising, since both complexity and degree of completeness are measures of the number of variable value appearances in a model. Whereas complexity simply counts the number of variable value appearances in a model, degree of completeness measures this number in proportion to the number of variable value appearances in the DGCS. So, the observed contrast implies that high contrapositive models tend to be inferred from less complex DGCSs than low contrapositive models in this experiment, which is possible if, for datasets generated from less complex DGCSs, it is more likely that at least one high contrapositive model is inferred. Thus, high contrapositive models make on average fewer causal relevance ascriptions than low contrapositive models, without sacrificing the proportion of the corresponding DGCS that they are able to recover. Finally, Figure 4 presents the numbers of search trials in which at least one low contrapositive model is returned (black bars) and in which at least one high contrapositive model is returned (grey bars). For many settings, only a small number of CNA outputs contain any high contrapositive models.

5.3 Discussion

The decrease in correctness for prevalence levels lower than consistency may be explained as follows: at lower prevalence, more conjuncts are needed to obtain a conjunction that reaches the required consistency threshold, leading to a higher number of causal relevance ascriptions, which makes it generally more likely that at least one of the causal relevance ascriptions is false. Indeed, Figure 2 illustrates that, as prevalence decreases below consistency, model complexity increases. That patterns in correctness and prevalence depend on consistency but are independent of coverage, may, first, be explained by the order of application of consistency and coverage in CNA’s model-building process: only after conjunctions reaching the consistency threshold have been built, are those conjunctions used for building disjunctions reaching the coverage

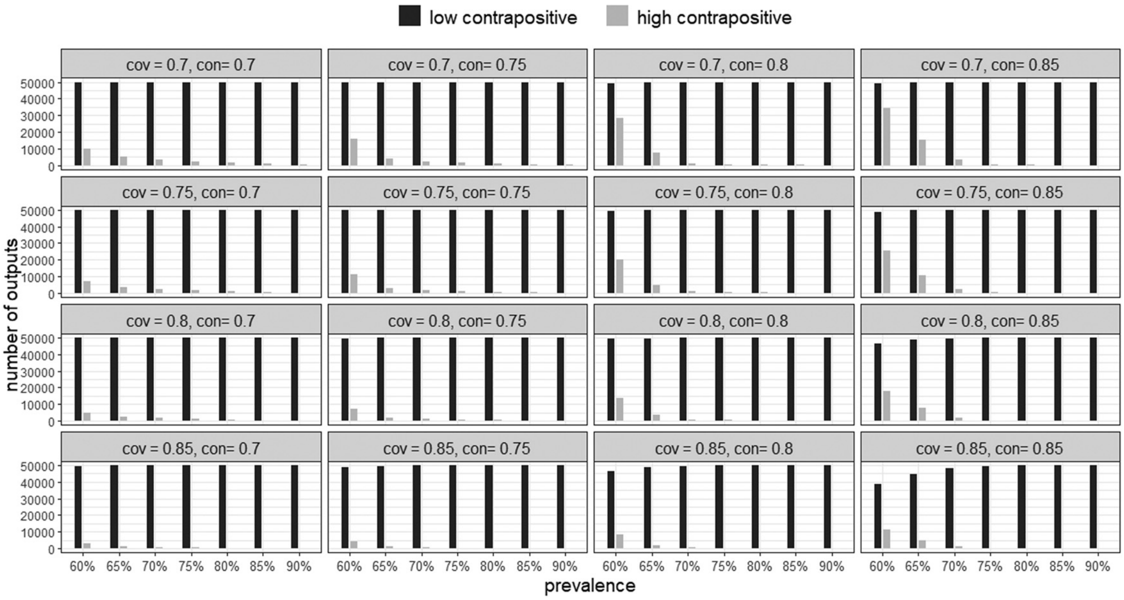

Figure 4: Number of recorded results.

threshold [16]. So, the range of possible models to be built is already severely limited by the consistency constraint before coverage comes into play, making coverage less influential than consistency in the model-building process. Second, as seen in Section 4, high consistency amounts to strong evidence for sufficiency when prevalence is low, and high coverage amounts to strong evidence for necessity when $|X|/N$ is low. The reliability of consistency is in this way directly linked to prevalence, whereas the reliability of coverage is linked to prevalence only indirectly, through the closeness of $|X|/N$ to prevalence as shown in Appendix A. This may also explain why patterns between correctness and prevalence are influenced more strongly by consistency than by coverage.

As argued for in Section 4, correctness decreases at prevalence levels higher than consistency because, for such settings, any variable value is likely to be accepted as sufficient for the outcome. Correctness of high contrapositive models does not decrease at prevalence levels higher than consistency because prevalence levels higher than the (contrapositive) consistency threshold do not increase the likelihood that models reach the *contrapositive* thresholds regardless of causal relevance. Nonetheless, the size of the difference in correctness between low and high contrapositive models, and the perfect or nearly perfect correctness scores for high contrapositive models when prevalence is at least as high as consistency, remain surprising. One additional explanation for the dramatic difference in correctness is a difference in complexity: for many settings, high contrapositive models make fewer than half as many causal relevance ascriptions as low contrapositive models. Still, as shown in Figure 3, high contrapositive models do not have a lower degree of completeness than low contrapositive models. High contrapositive models are less complex because, on average, less complex DGCSs lead to the discovery of high contrapositive models, not because high contrapositive models give a less complete picture of their corresponding DGCS.

5.4 Limitations

Before concluding, I discuss four limitations of the experiment, and one general potential limitation of contrapositive measures which can be evaluated in light of the results of the experiment. First, the noise in the experiment is generated by uniform random sampling from the cases incompatible with the DGCS. This sampling approach makes the experiment less representative for CCM analyses on datasets in which, due to systematic confounding, not all cases incompatible with the DGCS are contained with equal probability. It is expected that the performance of both the original measures, consistency and coverage, and the proposed measures, contrapositive consistency and contrapositive coverage, would be impaired by the presence of systematic confounding. However, since I do not claim that the proposed contrapositive measures would specifically mitigate the adverse effects of systematic confounding, and since there is no reason for suspecting that the performance of the proposed measures is impaired more by the presence of systematic confounding than the performance of the original measures, the use of uniformly distributed noise should not affect the interpretation of the results of the experiment.

Second, in this experiment, high prevalence levels are achieved by purposefully duplicating cases that have the desired outcome, amounting to sampling-induced prevalence variation. This is not the only way in which high prevalence can occur. For instance, prevalence may be high because the DGCS determines there to be many instances of the outcome, amounting to DGCS-induced prevalence variation. More research would be needed to find out whether the benefits of contrapositive measures for model quality are similar for DGCS-induced high prevalence. Still, as the prevalence of real-world datasets is typically sampling induced to some extent, the findings for sampling-induced high prevalence presented in this article are in any case informative. Third, even though the simulation experiment was conducted using CNA, the use of contrapositive measures for model evaluation should also be applicable to QCA, the most prominent CCM. Fourth, the experiment only includes crisp-set CCM analyses. Nevertheless, as shown in Appendix D, the findings of this article should be extendable to multi-value and fuzzy-set CCM analyses.

In addition, a possible disadvantage of relying on contrapositive consistency in high prevalence scenarios is that in such scenarios, only relatively few cases – those with y – are taken into account when calculating contrapositive consistency, and if some of those relatively few cases are affected by noise, the reliability of contrapositive consistency can be hurt substantially. However, the difference in correctness between low and high contrapositive models is expected to decrease with each correct model that is grouped with the low contrapositive models and each incorrect model that is grouped with the high contrapositive models, because correct and incorrect models tend to, respectively, increase and decrease the correctness of the group in which they are included. But, as the differences in correctness between low and high contrapositive models shown in Figure 1 remain very large, the results of the experiment suggest that the negative effects of this possible disadvantage are limited. So, this disadvantage is heavily outweighed by the advantages of taking into account contrapositive consistency and contrapositive coverage when prevalence is high. Furthermore, it should be noted that conventional coverage suffers from the same disadvantage when prevalence is low: in such scenarios, coverage is calculated based on relatively few cases. Analogously, consistency is calculated based on relatively few cases when $|X|/N$ is small and contrapositive coverage is calculated based on relatively few cases when $|X|/N$ is large.

6 Concluding remarks

The large increase in correctness and the preservation of the degree of completeness for high contrapositive models entail that, when high contrapositive models are returned, these must be the models of choice. Nevertheless, as shown in Figure 4, only a small number of trials produce any high contrapositive models, implying that there are also many correct models that do not meet the contrapositive thresholds. So, the results of the experiment should not be taken to imply that all low contrapositive models must be discarded. For models that do not meet the contrapositive thresholds, contrapositive consistency and contrapositive coverage should serve as extra model evaluation tools in addition to existing tools such as consistency and coverage, theoretical knowledge and case knowledge [4] (p. 172), and model robustness [40].

The explanations provided in Section 4 allow researchers to correctly interpret contrapositive consistency and contrapositive coverage and to appropriately use them in CCM model evaluation. Contrapositive consistency evaluates whether $X \rightarrow Y$ is satisfied among cases with y, and contrapositive coverage evaluates whether $X \leftarrow Y$ is satisfied among cases with x. Contrapositive consistency should be used as an additional measure for evaluating the sufficiency of full CCM antecedents and of conjunctions. In high prevalence datasets, even if contrapositive consistency does not meet the threshold chosen for consistency, reasonably high contrapositive consistency clearly amounts to stronger evidence in favour of sufficiency than lower contrapositive consistency. Contrapositive coverage should be used as an additional measure for evaluating the necessity of full CCM antecedents, and reasonably high contrapositive coverage clearly amounts to stronger evidence in favour of necessity than lower contrapositive coverage.

The findings of this article yield another recommendation for CCM practitioners: in order to obtain correct models for high-prevalence datasets, it is advisable to set the consistency threshold slightly above prevalence when conducting analyses using conventional consistency and coverage, as Figures 1 and 3 show that such consistency thresholds are most likely to produce correct models with a relatively high degree of completeness. Of course, this recommendation should be weighed against other considerations, such as the suspected proportion of noise in the dataset and the general requirement for reasonably high consistency. Furthermore, if one is willing to increase the risk of false causal relevance ascriptions in exchange for a possible gain in degree of completeness, as is customary in the so-called SI-approach to QCA [41] (p. 1874), increasing consistency further above prevalence may still be the preferred strategy.

Finally, the finding that contrapositive measures are able to select correct models but that only a small number of models produced by CCMs meet these thresholds suggests two approaches for further improving CCMs. The first approach is to develop more fine-grained ways of incorporating the evidence taken into account by contrapositive consistency and contrapositive coverage in model evaluation, as simply imposing contrapositive thresholds set equal to the conventional consistency and coverage thresholds was shown to exclude too many correct models. Lowering the contrapositive thresholds is not an optimal strategy for achieving this, because there is no principled method for deciding to what extent these thresholds should be lowered. A more promising way forward is to develop new measures that combine consistency with contrapositive consistency (for evaluating sufficiency) or that combine coverage with contrapositive coverage (for evaluating necessity).

For instance, the harmonic mean of consistency and contrapositive consistency takes into account more information than either of these measures alone. Moreover, a *weighted* harmonic mean of consistency and contrapositive consistency allows to assign more or less weight to consistency relative to contrapositive consistency depending on prevalence or $|X|/N$, which, as seen in Section 4, affect the relative importance of consistency versus contrapositive consistency in the evaluation of sufficiency. Analogously, a weighted harmonic mean of coverage and contrapositive coverage could suitably combine evidence captured by those two measures. Furthermore, developing prevalence-dependent weights for the penalties assigned to $X * y$ by consistency and contrapositive consistency and to $x * Y$ by coverage and contrapositive coverage (for one such proposal, see [24]) is a promising way of capturing evidence for sufficiency and necessity at varying prevalence levels.

A second approach for improving CCMs would benefit from the development of this first approach. A plausible explanation as to why only very few models in the experiment meet the contrapositive thresholds is that contrapositive measures were only used for model selection in this experiment and not for model building. But CCMs, as currently implemented, only use conventional consistency and coverage for model building and neglect some evidence that contrapositive measures are sensitive to. In order to obtain CCM models with higher contrapositive consistency and contrapositive coverage, this neglected evidence should be taken into account at the stage of model building already. This can be achieved by using a measure that combines consistency and contrapositive consistency in place of consistency and a measure that combines coverage and contrapositive coverage in place of coverage during model building.

Acknowledgements: I thank Michael Baumgartner for his invaluable feedback during the development of this article and the underlying research. In addition, I am grateful to Mathias Ambühl for providing R functions for applying the contrapositive measures. Furthermore, I benefited from discussions with Veli-Pekka Parkkinen, Martyna Swiatczak, and Torfinn Huvenes and from their feedback on earlier drafts. I am also indebted to three anonymous referees for their fruitful comments, and I thank the participants in the 2nd International Conference on Current Issues in Coincidence Analysis, the European PhD Network in Philosophy Seminar, and the University of Bergen Philosophy PhD Seminar for their constructive feedback.

Funding information: This work was funded by the Research Council of Norway (grant number 326215).

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated and analyzed during the current study can be reproduced with the R scripts available at https://github.com/Luna-De-Souter/Evaluating-Boolean-relationships-in-CCMs.

References

[1] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search (second edition). Cambridge: The MIT Press; 2000.
[2] Zhang J, Spirtes P. Detection of unfaithfulness and robust causal inference. Minds and Machines. 2008;18(2):239–71.
[3] Spirtes P, Zhang J. A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. Stat Sci. 2014;29(4):662–78.
[4] Oana IE, Schneider CQ, Thomann E. Qualitative Comparative Analysis using R: a beginner's guide. Methods for social inquiry. Cambridge: Cambridge University Press; 2021.
[5] Ragin CC. Fuzzy-set social science. Chicago: University of Chicago Press; 2000.
[6] Ragin CC. Set relations in social research: evaluating their consistency and coverage. Political Analysis. 2006;14(3):291–310.
[7] Tharwat A. Classification assessment methods. Appl Comput Informatics. 2021;17(1):168–92.
[8] Glass DH. Confirmation measures of association rule interestingness. Knowl Based Syst. 2013;44:65–77.
[9] Goertz G. Assessing the trivialness, relevance, and relative importance of necessary or sufficient conditions in social science. Stud Comp Int Dev. 2006;41(2):88–109.
[10] Schneider CQ, Wagemann C. Set-theoretic methods for the social sciences: a guide to Qualitative Comparative Analysis. Strategies for social inquiry. Cambridge: Cambridge University Press; 2012.
[11] Rothman KJ. Causes. Am J Epidemiol. 1976;104(6):587–92.
[12] Gerring J. Causation: a unified framework for the social sciences. J Theor Polit. 2005;17(2):163–98.
[13] Hart HLA, Honoré T. Causation in the law. Oxford: Oxford University Press; 1985.
[14] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics. 2009;10(6):392–404.
[15] Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet. 2002;70(2):461–71.
[16] Ambuehl M, Baumgartner M. cna: Causal modeling with Coincidence Analysis. 2023. R package version 3.5.2. https://CRAN.R-project.org/package=cna.
[17] Brambor T, Clark WR, Golder M. Understanding interaction models: improving empirical analyses. Political Analysis. 2006;14(1):63–82.
[18] Mackie JL. The cement of the universe: a study of causation. Oxford: Clarendon Press; 1974.
[19] Graßhoff G, May M. Causal regularities. In: Spohn W, Ledwig M, Esfeld M, editors. Current issues in causation. Paderborn: Mentis; 2001. p. 85–114.
[20] Baumgartner M, Falk C. Boolean difference-making: a modern regularity theory of causation. British J Philos Sci. 2023;74(1):171–97
[21] Baumgartner M, Ambühl M. Causal modeling with multi-value and fuzzy-set Coincidence Analysis. Political Sci Res Methods. 2020;8(3):526–42.
[22] Baumgartner M, Thiem A. Often trusted but never (properly) tested: evaluating Qualitative Comparative Analysis. Sociol Methods Res. 2020;49(2):279–311.
[23] Kuhn M, Johnson K. Measuring performance in classification models. In: Applied predictive modeling. New York: Springer; 2013. p. 247–73.
[24] Siblini W, Fréry J, He-Guelton L, Oblé F, Wang YQ. Master your metrics with calibration. In: Berthold MR, Feelders A, Krempl G, editors. Advances in intelligent data analysis XVIII. Cham: Springer International Publishing; 2020. p. 457–69.
[25] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Int J Machine Learn Technol. 2011;2(4):37–63.
[26] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
[27] Flach P, Kull M. Precision-recall-gain curves: PR analysis done right. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in neural information processing systems. vol. 28. New York: Curran Associates, Inc.; 2015.
[28] Koyejo OO, Natarajan N, Ravikumar PK, Dhillon IS. Consistent binary classification with generalized performance metrics. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Advances in neural information processing systems. vol. 27. New York: Curran Associates, Inc.; 2014.
[29] Hempel CG. Studies in the logic of confirmation (I.). Mind. 1945;54(213):1–26.
[30] Swiatczak MD. Towards a neo-configurational theory of intrinsic motivation. Motivation and Emotion. 2021;45(6):769–89.

[31] Csikszentmihalyi M. Beyond boredom and anxiety. San Francisco: Jossey-Bass Publishers; 1975.

[32] Seawright J. Testing for necessary and/or sufficient causation: which cases are relevant? Political Analysis. 2002;10(2):178–93.

[33] Clarke KA. The reverend and the ravens: comment on Seawright. Political Analysis. 2002;10(2):194–7.

[34] Braumoeller BF, Goertz G. Watching your posterior: comment on Seawright. Political Analysis. 2002;10(2):198–203.

[35] Haesebrouck T. Pitfalls in QCA's consistency measure. J Comparative Politics. 2015;8(2):65–80.

[36] Stoklasa J, Luukka P, Talášek T. Set-theoretic methodology using fuzzy sets in rule extraction and validation - consistency and coverage revisited. Inform Sci. 2017;412–413:154–73.

[37] Veri F. Coverage in fuzzy set Qualitative Comparative Analysis (fsQCA): a new fuzzy proposition for describing empirical relevance. Comparative Sociology 2018;17(2):133–158.

[38] Veri F. Aggregation bias and ambivalent cases: a new parameter of consistency to understand the significance of set-theoretic sufficiency in fsQCA. Comparative Sociology 2019;18(2):229–55.

[39] Mendel JM, Ragin CC. fsQCA: dialog between Jerry M. Mendel and Charles C. Ragin. USC-SIPI REPORT # 411. 2nd edition. 2012. https://ssrn.com/abstract=2517966.

[40] Parkkinen VP, Baumgartner M. Robustness and model selection in configurational causal modeling. Sociol Methods Res. 2023;52(1):176–208.

[41] Haesebrouck T, Thomann E. Introduction: causation, inferences, and solution types in Configurational Comparative Methods. Quality & Quantity. 2022;56:1867–88.

Appendix

A $|X|/N$ close to $|Y|/N$

This appendix shows that, if consistency and coverage are reasonably high, then $|X|/N$ is close to $|Y|/N$.

We have *consistency* = $|X * Y|/|X|$ and *coverage* = $|X * Y|/|Y|$.

From this, it follows that

$$\frac{\text{consistency}}{\text{coverage}} = \frac{|Y|}{|X|}.$$

$$\text{So}, \min\left(\frac{\text{consistency}}{\text{coverage}}\right) = \min\left(\frac{|Y|}{|X|}\right) \quad \text{and} \quad \max\left(\frac{\text{consistency}}{\text{coverage}}\right) = \max\left(\frac{|Y|}{|X|}\right).$$

This allows to determine the bounds of the ratio $\frac{|Y|}{|X|}$ based on the values of *consistency* and *coverage*:

$$\min\left(\frac{\text{consistency}}{\text{coverage}}\right) \le \frac{|Y|}{|X|} \le \max\left(\frac{\text{consistency}}{\text{coverage}}\right).$$

$\frac{\text{consistency}}{\text{coverage}}$ reaches its minimum when *consistency* is at its lowest and *coverage* is at its highest, and $\frac{\text{consistency}}{\text{coverage}}$ reaches its maximum when *consistency* is at its highest and *coverage* is at its lowest.

As an example, suppose that we set the minimum thresholds for *consistency* and *coverage* both at 0.7. *Consistency* and *coverage* can never be higher than 1, so 0.7 ≤ *consistency* ≤ 1 and 0.7 ≤ *coverage* ≤ 1. Then $\min\left(\frac{\text{consistency}}{\text{coverage}}\right) = 0.7/1 = 0.7$ and $\max\left(\frac{\text{consistency}}{\text{coverage}}\right) = 1/0.7 \approx 1.43$. So,

$$0.7 \le \frac{|Y|}{|X|} \lessapprox 1.43.$$

Because $|Y|/N$ and $|X|/N$ are equal when $|Y|/|X| = 1$ and close to each other when $|Y|/|X|$ is close to 1, these bounds imply that $|Y|/N$ is reasonably close to $|X|/N$. For example, if $|X|/N = 0.55$, $|Y|/N$ cannot be lower than $0.7 \cdot 0.55 \approx 0.39$ or higher than $1/0.7 \cdot 0.55 \approx 0.79$. So, even though $|X|/N$ is not guaranteed to be very close to $|Y|/N$, there can be no cases where $|X|/N$ is very high while $|Y|/N$ is very low. 0.7 is a rather low but still reasonable threshold, and as consistency and coverage thresholds increase, the maximum relative difference between $|X|/N$ and $|Y|/N$ decreases. So, if consistency and coverage are reasonably high, then $|X|/N$ is close to $|Y|/N$.

B Goertz's measure of nontrivialness from fuzzy to crisp sets

This appendix presents the fuzzy-set formulation of Goertz's measure of nontrivialness and shows a derivation of the crisp-set variant of this measure. The fuzzy-set measure is defined as follows [9] (p. 95):

$$\text{Goertz's measure of nontrivialness } (X \leftarrow Y) = \frac{1}{N_G} \cdot \sum_{i:\mathrm{X}_i \ge \mathrm{Y}_i} \frac{(1 - \mathrm{X}_i)}{(1 - \mathrm{Y}_i)}.$$

Note that the summation only includes cases with $\mathrm{X}_i \ge \mathrm{Y}_i$, because Goertz only defines his measure over cases that satisfy $X \leftarrow Y$. While in crisp-set data, $X \leftarrow Y$ is satisfied for all cases that are not of the type $x * Y$, in fuzzy-set data, it is, more generally, satisfied for cases with $\mathrm{X}_i \ge \mathrm{Y}_i$. Correspondingly, N_G stands for the number of cases with $\mathrm{X}_i \ge \mathrm{Y}_i$. As formulated here, Goertz's measure is not appropriate for use on crisp-set data, because it is undefined for cases with X = 1 and Y = 1: such cases would have a contribution of $\frac{(1-1)}{(1-1)} = \frac{0}{0}$. However, Goertz proposes to assign a contribution of 1 to these undefined cases to solve this issue (p. 108). I use this adaptation to formulate the crisp-set version of Goertz's measure.

As the summation only includes cases with $\mathrm{X}_i \geq \mathrm{Y}_i$, it only counts contributions from cases with $x * y$, cases with $X * Y$, and cases with $X * y$. Cases with $x * y$ have a contribution of $\frac{1-0}{1-0} = 1$, cases with $X * Y$ are defined by Goertz to have a contribution of 1, and cases with $X * y$ have a contribution of $\frac{1-1}{1-0} = 0$. So, the sum over the contributions of these cases is equal to $|x * y| \cdot 1 + |X * Y| \cdot 1 + |X * y| \cdot 0 = |x * y| + |X * Y|$. Since N_G is the number of cases with $\mathrm{X}_i \geq \mathrm{Y}_i$, $N_G = |x * y| + |X * Y| + |X * y|$. This gives the following crisp-set formulation:

$$\begin{aligned}\text{Goertz's measure of nontrivialness } (X \leftarrow Y) &= \frac{1}{N_G} \cdot \sum_{i:\mathrm{X}_i \geq \mathrm{Y}_i} \frac{(1-\mathrm{X}_i)}{(1-\mathrm{Y}_i)} \\ &= \frac{1}{N_G} \cdot (|x * y| + |X * Y|) \\ &= \frac{|x * y| + |X * Y|}{|x * y| + |X * Y| + |X * y|}.\end{aligned}$$

C Relevance of Necessity from fuzzy to crisp sets

This appendix presents the fuzzy-set formulation of Schneider and Wagemann's Relevance of Necessity measure and shows a derivation of the crisp-set variant of this measure. The fuzzy-set measure is defined as follows [10] (p. 236):

$$\text{Relevance of Necessity } = \frac{\sum_{i=1}^{N}(1-\mathrm{X}_i)}{\sum_{i=1}^{N}(1-\min(\mathrm{X}_i, \mathrm{Y}_i))}.$$

N stands for the number of cases in the dataset, so $N = |x * y| + |x * Y| + |X * y| + |X * Y|$. Furthermore, for crisp-set data, $\sum_{i=1}^{N}\mathrm{X}_i = |X|$, and fuzzy-set $\min(\mathrm{X}, \mathrm{Y})$ is equivalent to crisp-set $X * Y$. This allows for the following reformulation:

$$\begin{aligned}\text{Relevance of Necessity } &= \frac{\sum_{i=1}^{N}(1-\mathrm{X}_i)}{\sum_{i=1}^{N}(1-\min(\mathrm{X}_i, \mathrm{Y}_i))} \\ &= \frac{\sum_{i=1}^{N}1 - \sum_{i=1}^{N}\mathrm{X}_i}{\sum_{i=1}^{N}1 - \sum_{i=1}^{N}\min(\mathrm{X}_i, \mathrm{Y}_i)} = \frac{N - |X|}{N - |X * Y|} \\ &= \frac{|x|}{N - |X * Y|} = \frac{|x * y| + |x * Y|}{|x * y| + |x * Y| + |X * y|}.\end{aligned}$$

D Extension to multi-value and fuzzy set

D.1 Multi-value

This appendix shows that contrapositive measures can be applied to multi-value CCM models in the same way as they are applied to crisp-set models. Whereas crisp-set variables can take one of only two values for each case in the dataset, multi-value variables can take one of more than two values. Suppose that CCMs are used on a dataset with multi-value variables P, Q, R, S, and T, where P, Q, and R can take one of the values 1, 2, and 3, and S and T can take one of the values 1, 2, 3, and 4. Model (A1) is an example of a CCM model for this multi-value dataset.

$$\mathrm{P} = 3 * \mathrm{Q} = 1 + \mathrm{R} = 1 * \mathrm{S} = 4 \leftrightarrow \mathrm{T} = 2. \tag{A1}$$

Consistency can be formulated in the same way for this multi-value CCM model and dataset as for the crisp-set case:

$$\text{Consistency} = \frac{|X * Y|}{|X|}.$$

Here, $|X|$ stands for the number of cases in the dataset for which the antecedent of model (A1) is true, i.e. for which both P takes the value 3 and Q takes the value 1 or both R takes the value 1 and S takes the value 4. $|X * Y|$ denotes the number of cases in the dataset for which both the antecedent and the outcome of model (A1) are true, i.e. for which the antecedent is true and T takes the value 2.

Likewise, contrapositive consistency can be formulated in the same way for multi-value CCM models as for crisp-set CCM models:

$$\text{Contrapositive consistency} = \frac{|x * y|}{|y|}.$$

Here, $|y|$ represents the number of cases in the dataset for which the outcome of model (A1) is false. These are the cases for which T takes a value different from 2, i.e. value 1, value 3, or value 4. $|x * y|$ stands for the number of cases in the dataset in which both the antecedent and the outcome of model (A1) are false.

The formulation of coverage and contrapositive coverage for multi-value CCM models is analogous to the formulation of consistency and contrapositive consistency. So, contrapositive consistency and contrapositive coverage can be extended straight-forwardly from the crisp-set to the multi-value case.

D.2 Fuzzy set

This appendix shows how contrapositive measures can be applied to fuzzy-set CCM models. Fuzzy-set variables can take any real number between and including 0 and 1 as their value in each case of the dataset. Therefore, the classical Boolean operators conjunction, disjunction, negation, and implication are not directly applicable to fuzzy-set variables, introducing the need for fuzzy-logic variants of these Boolean operators. The fuzzy-logic operators standardly used in CCMs are as follows: conjunction $X * Y$ is defined as the minimum of X and Y, i.e. $\min(X, Y)$, disjunction $X + Y$ as the maximum of X and Y, i.e. $\max(X, Y)$, negation x is defined as $1 - X$, and implication $X \rightarrow Y$ as $X \leq Y$. This implies that consistency is formulated as follows for fuzzy-set data:

$$\text{Consistency} = \frac{\sum_{i=1}^{N} \min(X_i, Y_i)}{\sum_{i=1}^{N} X_i}.$$

As seen in Section 4.1, the justification for using contrapositive consistency in addition to consistency to evaluate $X \rightarrow Y$ in crisp-set CCMs is based on the law of contraposition, which entails that $X \rightarrow Y$ is logically equivalent to $y \rightarrow x$. This equivalence does also hold for the fuzzy-set case:

$$X \leq Y \Leftrightarrow X + 1 - X - Y \leq Y + 1 - X - Y \Leftrightarrow 1 - Y \leq 1 - X \Leftrightarrow y \leq x.$$

Hence, the justification for crisp-set contrapositive consistency is also applicable to its fuzzy-set variant. Fuzzy-set contrapositive consistency can be formulated analogously to crisp-set contrapositive consistency, yielding the following additional measure for evaluating $X \rightarrow Y$ in fuzzy-set CCMs:

$$\text{Contrapositive consistency} = \frac{\sum_{i=1}^{N} \min(x_i, y_i)}{\sum_{i=1}^{N} y_i}.$$

Note that the reformulations of crisp-set consistency and contrapositive consistency given at the end of Section 4.1 to show that both measures penalize $|X * y|$ are different for the fuzzy-set versions of these measures. Crisp-set consistency was reformulated as $1 - \frac{|X * y|}{|X|}$, showing that it penalizes $|X * y|$ in proportion to $|X|$. One might thus expect that fuzzy-set consistency penalizes $\sum_{i=1}^{N}\min(X_i, y_i)$ in proportion to $\sum_{i=1}^{N}X_i$. However, fuzzy-set consistency is reformulated as follows:

$$\text{Consistency} = \frac{\sum_{i=1}^{N}\min(X_i, Y_i)}{\sum_{i=1}^{N}X_i} = 1 - \frac{\sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N}X_i}.$$

Proof. The proof of this reformulation relies on the following equality:

$$\sum_{i=1}^{N}\min(X_i, Y_i) + \sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i) = \sum_{i=1}^{N}X_i.$$

By subtracting $\sum_{i=1}^{N}\min(X_i, y_i)$ from and adding $\sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)$ to both sides of this equality, we find that:

$$\sum_{i=1}^{N}\min(X_i, Y_i) = \sum_{i=1}^{N}X_i - \sum_{i=1}^{N}\min(X_i, y_i) + \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i).$$

This allows to reformulate fuzzy-set consistency as follows:

$$\begin{aligned}\text{Consistency} &= \frac{\sum_{i=1}^{N}\min(X_i, Y_i)}{\sum_{i=1}^{N}X_i}\\ &= \frac{\sum_{i=1}^{N}X_i - \sum_{i=1}^{N}\min(X_i, y_i) + \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N}X_i}\\ &= \frac{\sum_{i=1}^{N}X_i}{\sum_{i=1}^{N}X_i} - \frac{\sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N}X_i}\\ &= 1 - \frac{\sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N}X_i}.\end{aligned}$$

□

So, fuzzy-set consistency penalizes $\sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)$ in proportion to $\sum_{i=1}^{N}X_i$. Seeing that $\min(X_i, x_i, Y_i, y_i)$ is subtracted from the penalty gives insight in how cases with both $X_i < Y_i$ and $X_i < y_i$ influence consistency, as it renders explicit that these cases are rewarded rather than penalized by consistency: their contribution to the penalty is $\min(X_i, y_i) - \min(X_i, x_i, Y_i, y_i) = X_i - X_i = 0$ while they do increase the denominator of the penalty term (as long as they have $X_i > 0$) leading to a decrease in the total penalty (unless this penalty was already equal to 0). Many proposed fuzzy-set alternatives to consistency are a reaction to this and are developed in order to remove the reward [39] or even to introduce a penalty [36,38] for cases with both $X_i < Y_i$ and $X_i < y_i$.

Until now, most or all arguments against rewarding cases with both $X_i < Y_i$ and $X_i < y_i$ when evaluating sufficiency and necessity have been based on intuition rather than on logic, whereas the rationale for the contrapositive measures is based directly on the rule of contraposition from logic. The discussion on whether or not cases with both $X_i < Y_i$ and $X_i < y_i$ should be rewarded when evaluating Boolean sufficiency and necessity is outside the scope of this article. The important insight to be gained from the above reformulation is that consistency penalizes cases with $\min(X_i, y_i) > \min(X_i, x_i, Y_i, y_i)$.

Consistency penalizes these cases in proportion to $\sum_{i=1}^{N}X_i$. As seen in the following reformulation, contrapositive consistency penalizes exactly the same cases as those penalized by consistency, but contrapositive consistency penalizes these cases in proportion to $\sum_{i=1}^{N}y_i$ instead of in proportion to $\sum_{i=1}^{N}X_i$:

$$\begin{aligned}
\text{Contrapositive consistency} &= \frac{\sum_{i=1}^{N}\min(x_i, y_i)}{\sum_{i=1}^{N} y_i} \\
&= \frac{\sum_{i=1}^{N} y_i - \sum_{i=1}^{N}\min(X_i, y_i) + \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N} y_i} \\
&= \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i} - \frac{\sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N} y_i} \\
&= 1 - \frac{\sum_{i=1}^{N}\min(X_i, y_i) - \sum_{i=1}^{N}\min(X_i, x_i, Y_i, y_i)}{\sum_{i=1}^{N} y_i}.
\end{aligned}$$

The formulation of coverage and contrapositive coverage for fuzzy-set CCM models is analogous to the fuzzy-set formulations of consistency and contrapositive consistency. So, contrapositive consistency and contrapositive coverage can be extended from the crisp-set to the fuzzy-set case.