Article

Sociological Methods & Research
1-20

DOI: 10.1177/0049124119882460
journals.sagepub.com/home/smr
SAGE

The Double Bind of Qualitative Comparative Analysis

Vincent Arel-Bundock[1]

Abstract

Qualitative comparative analysis (QCA) is an influential methodological approach motivated by set theory and boolean logic. QCA proponents have developed algorithms to analyze quantitative data, in a bid to uncover necessary and sufficient conditions where causal relationships are complex, conditional, or asymmetric. This article uses computer simulations to show that researchers in the QCA tradition face a vexing double bind. On the one hand, QCA algorithms often require large data sets in order to recover an accurate causal model, even if that model is relatively simple. On the other hand, as data sets increase in size, it becomes harder to guarantee data integrity, and QCA algorithms can be highly sensitive to measurement error, data entry mistakes, or misclassification.

Keywords

qualitative comparative analysis, Monte Carlo simulation, measurement error, sample size, research methods, configurational diversity

Qualitative comparative analysis (QCA) is an influential methodological approach motivated by set theory and boolean logic. It was originally developed in the 1980s to integrate and formalize tools for the comparative

[1] Université de Montréal, Quebec, Canada

Corresponding Author:
Vincent Arel-Bundock, Université de Montréal, 3150 Jean Brillant, Montreal, Quebec, Canada H3T 1N8.
Email: vincent.arel-bundock@umontreal.ca

analysis of macro-historical social processes (Ragin 1987 2009). Since then, QCA has been deployed in thousands of published articles, in a wide variety of domains including political science, sociology, management, public health, environmental science, education, and criminology (Roig-Tierno, Gonzalez-Cruz, and Llopis-Martinez 2017).

QCA is perhaps best understood as a holistic "approach," which aims to identify interwoven "conditions of occurrence" (Berg-Schlosser et al., 2009). It stresses the need to integrate qualitative analysis and more formal quantitative methods. To that end, QCA offers a set of tools and algorithms for the analysis of quantitative data (Berg-Schlosser et al., 2009). Proponents argue that those algorithms can uncover necessary and sufficient conditions in complex, conditional, and asymmetric causal relationships. They also suggest that QCA algorithms perform well in the analysis of small-to-intermediate data sets (e.g., 10 to 50 observations), where multiple regression may be less useful (Berg-Schlosser et al., 2009).

Several researchers have used computer simulations to probe the robustness and accuracy of the algorithmic tools developed in the QCA tradition (e.g., Baumgartner and Ambühl 2019; Hug 2013; Krogslund, Choi, and Poertner 2014; Lucas and Szatrowski 2014). These authors take the value of *qualitative* analysis as given and assess how well the QCA tool kit performs in *quantitative* data analysis. They manipulate features of the data or estimation procedure in experimental fashion to identify the conditions under which QCA routines yield correct answers.

This article builds on prior simulation-based work to highlight a vexing double bind. In most real-life applications of crisp set QCA, some combinations of explanators are more likely to occur than others. In such cases, the performance of available algorithms can suffer,[1] and the sample sizes that are needed to produce satisfactory results become much larger. When a sample increases in size, it becomes harder for researchers to guarantee data integrity. As this article shows, measurement error, data entry mistakes, classification problems, or typological ambiguity can have deleterious effects on QCA solution quality.[2] Thus, researchers in the QCA tradition often face a choice between studying small data sets with limited configurational diversity or large data sets with measurement error. In both contexts, QCA algorithms can fail to recover complete or truthful causal models.

In support of this argument, I present the results of extensive Monte Carlo experiments. These experiments yield three main conclusions. First, the sample size required for credible inference using QCA is much larger than is usually acknowledged in the literature. In the typical case where some combinations of explanators are more likely to occur than others, analysts

have no guarantee that crisp set QCA will recover a simple four-variable model, even if the number of observations exceeds 300.

Second, a very small amount of measurement error can severely degrade the quality of QCA solutions. For instance, in data sets with 49 "good" observations and a single "bad" one, about 30 percent of the causal claims produced by a QCA algorithm are strictly incorrect.[3] Three mismeasured observations make the proportion of false causal claims jump to about 60 percent.[4] Even if the sample exceeds 300 observations, one would need exceptional data quality or configurational diversity to ensure that the routine produces less than 5 percent of incorrect claims.[5]

Third, the tuning parameters that users select for their QCA algorithms can have a major effect on inference. For example, analyzing a data set with 300 observations, a single error, and the *default* consistency threshold of a popular QCA routine produces nearly 40 percent of false causal claims. Introducing three mistakes in a data set of 300 observations makes the share of false causal claims cross the 70 percent bar.

Beyond these substantive insights, this article also makes three methodological contributions to the simulation-based literature on QCA. First, I propose a flexible mechanism to manipulate configurational diversity in Monte Carlo simulations. This mechanism draws an explicit link between diversity and a quantity of great practical interest to empiricists: sample size. The results described below thus offer strong intuition about the sample sizes which ensure configurational diversity and about the number of observations that QCA users should make in order to draw credible inference about complex conditional processes.

Second, to test the effect of sample size and measurement error on QCA solutions, I introduce two new formal criteria: *wrongness* and *completeness*. Those two criteria measure the extent to which QCA solutions are compatible with the data generating process. In addition, I evaluate the performance of QCA algorithms by adapting a measure of classification accuracy which is standard in the statistics and machine learning literatures: root mean squared error (RMSE). Taken together, those three criteria offer a more comprehensive and fine-grained view of QCA performance than is typical in simulation studies.

Third, the results reported in this article improve on prior work by taking to heart two major critiques of QCA simulation studies. To begin, my simulations consider that even if a QCA solution does not match the *complete* true model, that solution could still be a correctness-preserving *submodel* of the truth (Baumgartner and Thiem 2017b). In addition, my simulations take into account the presence of model ambiguities, that is, the possibility that several

boolean causal models could be consistent with the observed data (Baumgartner and Thiem 2017a; Rohlfing 2015). Since very few (if any) simulation studies answer both of these challenges, the results presented here constitute an important step forward in terms of offering a credible assessment of QCA performance.

Taken together, the results of this study suggest that QCA analysts face a double bind: When configurational diversity is limited, QCA algorithms need large samples to recover complex boolean models of conditional causal processes. As we increase the size of our samples, however, measurement error risks being introduced, and this can negatively affect the performance of our algorithms. This conclusion reinforces the crucial role that qualitative analysis must play in QCA, as it can help guard against measurement error and limit the inferential risk posed by incorrect algorithmic solutions.

A Double Bind

In its original incarnation, QCA was developed for the analysis of substantive problems where a relatively small number of cases could be compared. Later, QCA algorithms were used to analyze much larger data sets, with observations numbering in the thousands. Yet, most QCA methodologists still tout its advantages in the analysis of data sets of small-to-intermediate size, and this remains the predominant use case.

To illustrate, I collected data on the number of cases studied in 199 peer-reviewed articles published between 2016 and 2019 using QCA techniques. Twenty-five percent of those studies considered fewer than 25 cases, and 50 percent leveraged information on less than 63 cases. Figure 1 shows the full distribution of sample sizes across the 199 studies.[6] Based on these data, it seems fair to say that the performance of QCA in relatively small samples remains critical.

One important practical challenge is that if the number of observations is limited, configurational diversity may also be limited. As all practitioners know, some combinations of explanatory factors are usually more likely to occur than others in real-life data sets. Missing data can also limit the types of cases that one can consider for analysis. Often, some combinations of causal factors are not observed at all. When configurational diversity is limited, QCA algorithms may fail to recover the full causal model.

The most obvious way to improve configurational diversity is to collect more data. However, increasing the sample size comes with an important drawback. As more units are observed, it becomes more difficult to guarantee

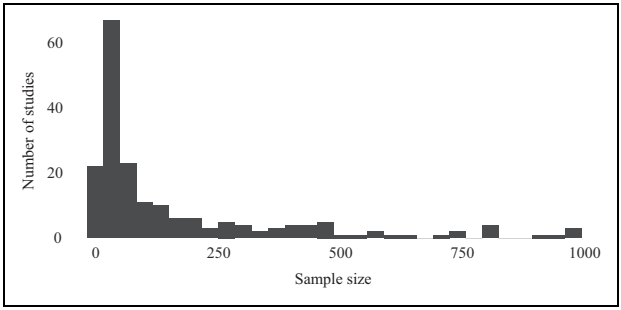

Figure 1. Number of cases considered in 199 qualitative comparative analysis applications (2016–2018).

the integrity of our data sets, that is, it becomes more difficult to limit measurement error.

Measurement error comes in different flavors. In its simplest form, it could be the result of a simple data entry mistake, when a researcher (or research assistant) fails to record the correct value for a given unit of observation. It can also arise at the data processing stage, as in the case of Reinhart and Rogoff (2010), who mistakenly excluded five important countries from their analysis of public debt and economic growth (Thomas, Michael, and Robert 2014). Measurement error can also be introduced before analysts even get a hold of their data. For example, in 2014, the gross domestic product of Nigeria doubled overnight, following an overdue recalculation by the country's National Bureau of Statistics (Mezue 2014). Similarly, Linsi and Mügge (2019) show that international economic statistics are systematically biased, and Kerner, Jerven, and Beatty (2017) suggest that national statistical agencies could be engaging in "aid-seeking data management." Of course, measurement error could also be related to ambiguities of a more conceptual nature. If a researcher's theory is imprecise, observable units could be misclassified or mischaracterized, which has important implications for the analysis of sufficient and necessary conditions.

The goal of this article is to measure analysts' ability to escape the double bind posed by limited configurational diversity and measurement error. To put this exercise in proper context, it is useful to note that the task that QCA routines set for themselves is a very difficult one because their search target is more complex than that of typical regression analyses. It is obviously more difficult to recover a complex boolean ordering than a simple conditional expectation. The fact that QCA researchers often try to recover complex

models from limited data compounds the difficulty. Under these circumstances, it should not be surprising to find that measurement error and sample size have some effect on the performance of QCA algorithms. As always, what matters is that practitioners know where a tool can be expected to perform well and that they be able to assess the degree of scientific uncertainty that remains postanalysis.[7]

In that constructive spirit, the next section introduces three fine-grained evaluation criteria which can be used to evaluate the performance of QCA algorithms. These formal criteria are then put to use in extensive Monte Carlo simulations. The results offer some of the most credible estimates to date of the performance of crisp set QCA in a range of realistic conditions.

Evaluation Criteria

To evaluate the fitness of QCA solutions, I adopt two complementary strategies. The first approach is most common in the field of qualitative comparative analysis, where complex combinations of causal factors are the main object of interest. The second approach is more common in the fields of statistics and machine learning, where classification (or predictive) accuracy is often a central concern.

Data Generating Process

The goal of QCA algorithms is typically to find a complex boolean ordering over a set of causal factors. Therefore, when evaluating the performance of a QCA solution, it makes sense for simulation studies to test whether that solution is logically compatible with the true data generating process. To do this, I propose two new fitness criteria: *wrongness* and *completeness*.

The wrongness and completeness concepts build on work by Baumgartner and Thiem (2017b), two leading QCA methodologists who formulated sharp rebukes of prior simulation-based work in this field. In a recent *Sociological Methods & Research* article, these authors claim that although QCA is "often trusted," it is "never (properly) tested." The crux of their critique is simple, yet powerful: In the face of causal complexity, QCA may not always retrieve the full model of the truth, but it can nevertheless identify "submodels" that are consistent with the truth.

For example, consider an event Z which occurs if at least one of three conditions obtains: $(\neg A \wedge B) \vee (B \wedge \neg C) \vee D$. Typically, simulation-based tests check if QCA recovers the full causal model and count as a mistake any result that does not exactly match the truth. However, as Baumgartner and

Thiem point out, a simpler solution like $(B \wedge \neg C) \vee D$ should not necessarily be considered a "mistake," since that submodel remains logically compatible with the truth. More generally, the authors argue that evaluating the fitness of QCA solutions requires us to consider whether those solutions are "correctness-preserving" submodels of the truth.

Intuitively, a correctness-preserving submodel is a set of conditions which are simpler than the reference model, while remaining logically consistent with that benchmark. In other words, even if a submodel does not capture the full truth, the factors that it identifies as relevant still lie on the causal path to the outcome variable.[8] Since most existing simulation-based tests do not take submodels into account, they could overstate the practical problems of QCA.

To improve the credibility of simulation results, I introduce two fine-grained and complementary measures of solution fitness:

1. *Wrongness*: A QCA solution is "wrong" if at least one of its submodels is not a submodel of the truth. I measure the level of wrongness by counting the proportion of solution submodels that are not submodels of the truth.
2. *Completeness*: A QCA solution is "complete" if all the submodels of the truth are submodels of the QCA solution. I measure the level of completeness by counting the proportion of submodels of the truth that are also submodels of the solution.

Roughly speaking, we can think of a QCA solution as allowing analysts to make a certain number of claims about causally relevant variables. *Wrongness* measures the share of causal claims that are ostensibly supported by the QCA solution but which are in fact erroneous. *Completeness* measures the share of possible true claims that our QCA solution captures.[9]

Classification Accuracy

Researchers who work in the QCA tradition do not usually report measures of classification accuracy, since classification is not always considered to be an explicit goal of QCA analysis. Nevertheless, it can be useful to approach the problem from this perspective, since it allows us to build a bridge between QCA and conventional statistics; classification-based fitness criteria can easily be understood and interpreted by researchers working in both traditions.

One natural way to measure the performance of QCA solutions in terms of classification power is to test whether a given QCA solution correctly

classifies units with different combinations of explanatory variables. For example, if the true model of the world is

$$Z \Leftrightarrow (\neg A \wedge B) \vee (B \wedge \neg C) \vee D,$$

and if an individual has traits $A = 0, B = 1$, then we expect $Z = 1$. If a QCA algorithm produces the following (incomplete) solution candidate:

$$Z \Leftrightarrow (\neg A \wedge B) \vee D,$$

we would expect that same individual to exhibit $Z = 1$ as well. Even if the candidate solution is incomplete, it makes the correct "classification." To measure the performance of a QCA solution, we can simply calculate its RMSE over the 16 possible combinations of A, B, C, D:

$$\text{RMSE} = \sqrt{\frac{1}{16}\sum_{i=1}^{16}(Z_i - \hat{Z}_i)^2},$$

where Z_i is the true value of Z for the ith configuration, and $\hat{Z}_i$ is the value of Z that is consistent with the QCA solution candidate.

Simulation Design

To test the performance of QCA, I take three steps: (1) simulate hundreds of thousands of data sets governed by a known causal model; (2) estimate QCA models to extract solutions from each data set;[10] and (3) measure the wrongness, completeness, and RMSE of each solution. Figure 2 gives a more detailed description of this Monte Carlo design.

For continuity and simplicity, I consider the same data generating process as in Baumgartner and Thiem (2017b). Each data set includes five binary variables which conform to this law: $Z \Leftrightarrow (\neg A \wedge B) \vee (B \wedge \neg C) \vee D$.

Sample Size and Configurational Diversity

As mentioned above, it seems reasonable to expect that configurational diversity will affect the quality of solutions proposed by a QCA algorithm. Consider the same case as above, where five binary variables are related by this law: $Z \Leftrightarrow (\neg A \wedge B) \vee (B \wedge \neg C) \vee D$. It is easy to show that 16 different combinations of A, B, C, D, Z satisfy this causal model. If only a small fraction of those 16 configurations are actually observed empirically, QCA algorithms can hardly be blamed for failing to retrieve the complete causal model, since there would not be enough information to do so. Put simply,

1. Simulate
   (a) Choose a true causal process: $Z \Leftrightarrow (\neg A \wedge B) \vee (B \wedge \neg C) \vee D$, where A, B, C, D, Z are binary variables.
   (b) Identify the 16 configurations of A, B, C, D, Z that conform to the truth.
   (c) Identify the 16 configurations of A, B, C, D, Z that do not conform to the truth.
   (d) Take n observations from the set of truth-conforming configurations by drawing with replacement using constant (or Log-Normal) sampling weights.
   (e) Take k observations from the set of truth-breaking configurations by drawing with replacement using constant sampling weights.
   (f) Combine the n truth-conforming and the k truth-breaking observations into a single dataset.
2. Estimate
   (a) Build a truth table and compute a parsimonious QCA solution using the `truthTable` and `eQMC` functions from the `QCApro` library for `R`.
3. Measure
   (a) Calculate the wrongness, completeness, and RMSE of each solution.

Repeat steps one, two, and three 5000 times.

Figure 2. Monte Carlo simulation design.

the completeness and wrongness of QCA solutions surely depend on the diversity of our data sets.

Unfortunately, even if drawing a link between configurational diversity and QCA solution quality is commonplace, that link is of limited use to practitioners. Unless all the possible combinations of explanators are actually observed, data analysts will *never* be certain that their data are diverse enough because this would require a priori knowledge of the true causal model.[11] In this article, I use explicit distributional assumptions to tie the concept of configurational diversity to the size of a data set. This is an important contribution because it can help analysts develop intuition about the sample size that they need to draw credible inference using QCA.

In an ideal world, each truthful configuration of explanators would occur with equal frequency in our data sets. This guarantees a certain level of configurational diversity, even with relatively few data points. As every analyst knows, however, some configurations of variables are much more likely to occur than others in real-life data sets.

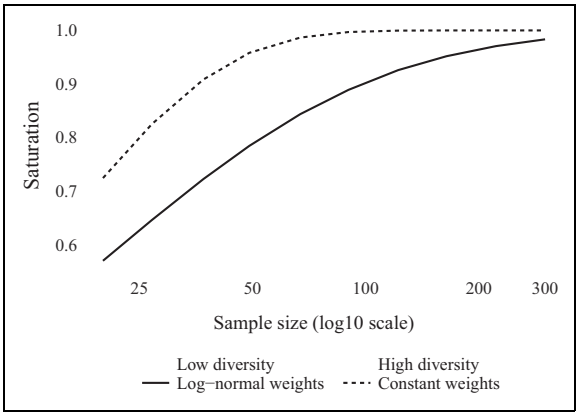

Figure 3. On average, what share of the 16 truthful configurations of A, B, C, D, Z are observed in simulated data sets?

To operationalize this intuition, I create simulated data sets with sample sizes ranging from 20 to 300. In some simulations, I assign a constant sampling weight to each true configuration of A, B, C, D, Z; this ensures that each of them will be equally likely to be observed. In other simulations, I assign sampling weights drawn from a log-normal distribution ($\mu = 0; \sigma = 1$) to each true configuration of A, B, C, D, Z; this ensures that some of them will be more likely to be observed than others.[12]

Figure 3 illustrates the distinction. In the unrealistic best-case scenario, where each truthful configuration has the same chance of occurring (constant sampling weights), we need a sample size of just under 100 observations to reach full configurational diversity. When some configurations are more likely than others (log-normal sampling weights), even a large sample size (e.g., $n = 300$) does not ensure saturation. Importantly, these results refer to data sets with only four binary explanators. Reaching full configurational diversity will obviously require larger data sets when the number of explanators increases.

Measurement Error

To study the effect of measurement error, I introduce 0, 1, 2, or 3 randomly selected "bad" observations in each data set. An observation is considered "bad" if its combination of A, B, C, D, Z violates the true causal model.[13] Introducing those bad observations allows us to assess the effect of

measurement error, classification problems, typological ambiguity, and data entry mistakes on QCA solution quality.

Consistency and Overfitting

Ragin (2006:292) introduced the concept of "set theoretic consistency" to QCA analysis, defining it as "the degree to which the cases sharing a given condition or combination of conditions [. . .] agree in displaying the outcome in question." Many modern QCA software routines allow users to define a consistency threshold, which relaxes the need for every case to conform exactly to a single causal process. It is well-known that QCA can overfit data when the consistency thresholds of algorithms are set too high (Schneider and Wagemann 2012). Thus, when analysts expect some measurement error, it makes sense for them to choose a lower threshold. As Rohlfing (2015) argues, ignoring overfitting in simulation studies could produce misleading results.

To explore the importance of user-selected settings for QCA performance, I report two sets of results: one with the consistency threshold set to 1.0 and the other with the consistency set to 0.75. The 1.0 threshold is important because it could be selected by applied researchers who mistakenly believe that their data are error free. It is also the default threshold used by several QCA software routines. Even if applied researchers have access to clear guidelines and best practices, we know that software defaults remain extremely important because they have strong behavioral effects. The 0.75 threshold is also important because it reduces the likelihood of overfitting and because it has become a focal point of sorts in both applied work and simulation-based studies (Baumgartner and Ambühl, 2019).

Model Ambiguities

Wrongness, completeness, and RMSE are solution-level criteria. However, QCA algorithms often propose a "model space" composed of several candidate solutions rather than a single solution. For instance, in the thousands of simulations conducted for this article, QCApro produced a unique solution in 77 percent of cases, two candidates in 13 percent of cases, and three in 2 percent of cases. At the other end of the spectrum, one data set yielded a model space with 72 candidates.

The ability to offer several solution candidates is an important feature of QCA routines. Indeed, it has long been recognized that several logical models can be compatible with the same set of configurational data (Simon 1954; Sprites, Glymour, and Scheines 2000). When empirical data underdetermine

their own causal modeling, the proper thing to do is to display epistemological humility and to propose the full range of data-consistent model candidates.

In a recent contribution, Baumgartner and Thiem (2017a) argue that model ambiguities are widespread in real-life applications of QCA and that failure to consider the full range of data-consistent models can lead researchers astray. Baumgartner and Thiem (2017b) also point out that when a QCA routine produces several model candidates, those candidates should be interpreted disjunctively. A QCA routine succeeds if it includes the correct model as one of the candidates in its model space. Clearly, model ambiguity needs to be considered in any simulation exercise (Rohlfing 2015).

To move from the solution level of analysis to the model space level of analysis, we thus ask the following question: How good is the *best* solution in each model space? More specifically, to measure the wrongness of a QCA model space, we take the minimum level of wrongness of all the solution candidates. To measure completeness, we take the maximum level of completeness.[14] To measure the RMSE, we take the minimum RMSE.

Focusing on the best available solution is generous to the QCA algorithm, it is logically correct, and it is consistent with prior methodological work. One potential downside of that approach is that it could yield simulation results which are less directly relevant to applied researchers. Indeed, practitioners will never know which of the proposed candidates is the best one in any given application. Focusing on maximum completeness and minimum wrongness could also favor less informative methods which generate a lot of candidate solutions. For instance, a model space with 50 candidates could have perfect scores on all criteria, even if it includes 49 wildly incorrect solutions. Despite these trade-offs, the proposed treatment of model ambiguities seems like the most principled approach.

For the rest of the analysis, we will only consider the best available QCA solution in any given model space. In the discussion of results, a claim such as "*X* percent of the causal claims are correct" will be shorthand for "*X* percent of the causal claims compatible with the best QCA solution in the model space are correct."

Results

Figure 4 reports the average level of wrongness of QCA solutions.[15] The most striking finding is that when using the default consistency threshold of 1.00, a *single* truth-breaking observation suffices to lead QCA routines astray in a large proportion of data sets.[16] In a sample of 300 observations with a single bad case, about 38 percent of the causal claims which are ostensibly

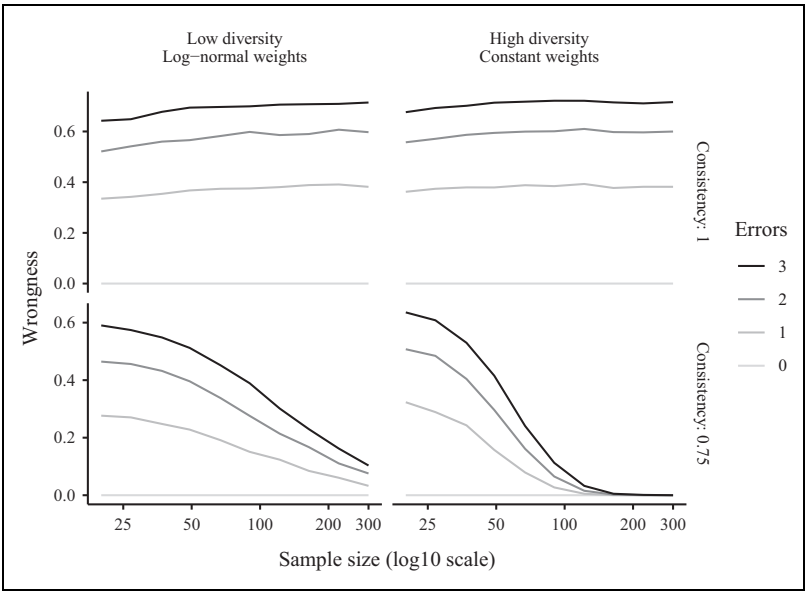

Figure 4. Average wrongness of parsimonious qualitative comparative analysis solutions in Monte Carlo simulations.

supported by QCA turn out to be false. Introducing three bad observations (of 300) makes the proportion of false causal claims jump above 70 percent. Importantly, the wrongness does not improve as we increase the number of "good" observations in the data set.

Setting the consistency threshold to 0.75 improves matters considerably but only in large data sets. Moreover, the level of wrongness remains high in most realistic settings. For instance, when the sample includes 20 good units and 1 erroneous observation, QCA produces about 30 percent of incorrect causal claims. When there are three erroneous observations, about 60 percent of causal claims are incompatible with the truth. When the sample size includes 50 units with one error, about 20 percent of solution submodels are wrong. In short, to ensure that a QCA algorithm will produce an acceptable number of false claims, researchers need to analyze a large data set of exceptional quality using the correct tuning parameters.

Figure 5 reports the average level of completeness of QCA solutions. The first conclusion to draw is that the sample size required to recover a complete boolean model is much larger than is typically acknowledged in the literature.

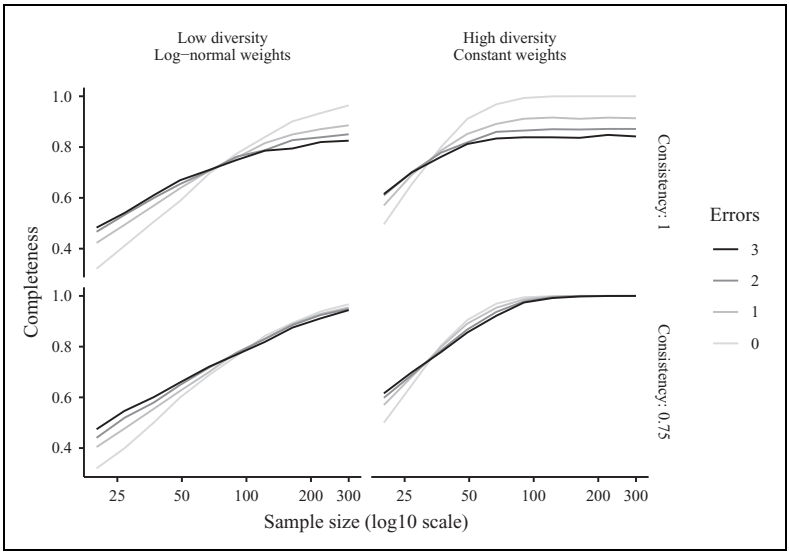

Figure 5. Average completeness of parsimonious qualitative comparative analysis solutions in Monte Carlo simulations.

Unless a sample includes over 70 observations, QCA routines are unlikely to recover the full causal process, even in a simple four-variables case. Moreover, when some configurations are more likely to be observed than others, the rate of convergence toward 100 percent completeness slows down considerably. When some configurations are observed more frequently than others, researchers will typically need over 300 observations to recover a complete model.[17]

The second result to notice is that, in the presence of measurement error, the consistency threshold has an important effect on performance: The top right panel shows a substantial decrease in completeness when errors are introduced, but not the bottom right panel. This reinforces the idea that users should be extremely careful when choosing the tuning parameters of their preferred QCA software. Given that measurement error is ubiquitous in social science and considering the powerful effect that software defaults can have on user behavior, these results also suggest that QCA software developers must seriously consider the default parameters of their routines.

These results also show that, in most real-life applications, limited configurational diversity forces researchers to use large data sets if they hope to recover complete QCA solutions. Unfortunately, as we increase the size of data sets, it becomes harder to guarantee data integrity, and

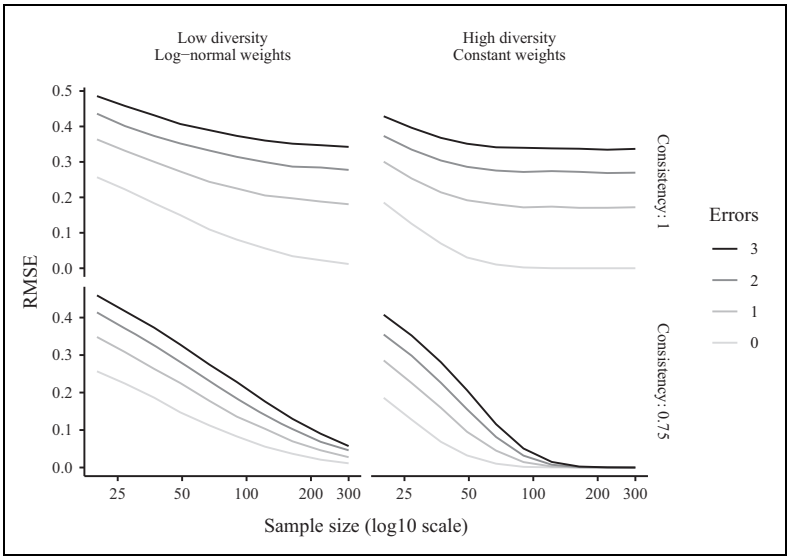

Figure 6. Root mean squared error of parsimonious qualitative comparative analysis solutions calculated in Monte Carlo simulations.

QCA routines can produce false causal claims in the presence of measurement error.

So far, we have assessed whether QCA solutions are compatible with the data generating process used to simulate data. A complementary approach is to measure the classification accuracy of QCA solutions. Figure 6 reports the average RMSE of the solutions computed for the same simulated data we used thus far. The results suggest that measurement error, sample size, configurational diversity, and tuning parameters can have important effects on performance.

Specifically, Figure 6 shows that introducing a single error can sometimes increase the RMSE of a QCA model space by 100 percent or more. With the exception of cases where the sample is very large and configurational diversity is high, introducing a single error increases RMSE by at least 50 percent. Introducing three errors can increase the RMSE by a factor of three or four.

Figure 6 also offers good news for QCA advocates: RMSE trends downward as the sample size increases, and this downward trend is accelerated when users select appropriate tuning parameters. To minimize classification error, the main challenge is, again, to increase the sample size while preserving data integrity.

Conclusion

This article introduced two new formal criteria, and adapted a third one, in order to evaluate the performance of QCA in simulated data: wrongness, completeness, and RMSE. These criteria have intuitive interpretations and allow us to take into account two major critiques of prior simulation-based efforts to evaluate QCA. Specifically, the simulations described above take into account the possibility of model ambiguities, and the importance of correctness-preserving submodels (Baumgartner and Thiem 2017b; Rohlfing 2015).

Based on extensive Monte Carlo simulations, I conclude that crisp set QCA algorithms can be very sensitive to measurement error. In principle, analysts could circumvent this problem by building small data sets that are completely error free. Unfortunately, the results presented here also suggest that QCA requires rather large data sets in order to work effectively.

Given the vast potential for human error at all stages of the scientific process—from the development of a typology, to measurement, classification, and data entry—it seems likely that most large (or even medium) sized data sets will contain *some* measurement error. Unless data analysts can guarantee that their data sets are completely error free, and unless they can pick optimal tuning parameters for QCA algorithms, prudence in interpretation is warranted.

Debates on the value of QCA often revolve around the distinction between deterministic and probabilistic types of arguments. But one does not need to make strong ontological commitments to conclude that the extreme sensitivity of QCA algorithms to measurement error is a major problem; all one needs to do is acknowledge that the social scientists who build data sets are human and fallible. The field of statistics was transformed in the 1980s by the development of robust high breakdown point estimators, which could yield reliable inference in the presence of outliers (Huber, 1981). If QCA proponents want the approach to flourish, they will likewise need to innovate and offer algorithms that are more robust to mild departures from the ideal world.

Acknowledgment

The author thanks Damien Bol, Gabrielle Péloquin-Skulski, Bear Braumoeller, Christopher Winship, and two excellent reviewers.

Declaration of Conflicting Interests

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

Funding

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

ORCID iD

Vincent Arel-Bundock https://orcid.org/0000-0003-2042-7063

Notes

1. On limited diversity and qualitative comparative analysis (QCA), see Seawright (2014) and Baumgartner and Thiem (2017b).
2. This article builds on previous work by Hug (2013) and Baumgartner and Ambühl (2019). Hug introduces random variation in several existing data sets to assess the effect of noise on QCA solutions. The author does not take into account model ambiguities or correctness-preserving submodels, and the fitness criteria that he uses to evaluate QCA solutions are too coarse to allow the rich and intuitive interpretation offered below. Baumgartner and Ambühl (2019) focus on a comparison of QCA to multivalue coincidence analysis. Their simulations allow some measurement error, but the data generating process ensures that the user-supplied consistency threshold is always satisfied. Since applied researchers will never know exactly what specific threshold to use in any given application, that condition is likely to be violated in empirical work. In addition, the authors use a very different (binary) fitness criteria: the QCA output passes a test if it includes a single correctness-preserving model that is a submodel of the truth. As such, their results are best viewed as complementary rather than as comparable to mine.
3. Here, the expression "causal claims" refers to the list of submodels which are logically compatible with the best QCA solution proposed by the algorithm for a given data set.
4. The 30 percent and 60 percent figures represent the average wrongness in simulations with limited configurational diversity and different consistency thresholds. Formal definitions of these terms are given below.
5. At most 1 mistake of the 300, or perfect equiprobability of configurations. These figures understate the severity of the problem in cases where a QCA algorithm proposes several model candidates instead of a single QCA solution. As I explain below, the simulation results described herein only consider the best solution proposed by the QCA routine for any particular data set. Unfortunately, applied researchers will never know which model is best in any given application.
6. This histogram does not distinguish between applications of crisp set, fuzzy set, or multivalue QCA. Six cases with sample sizes over 1,000 are omitted from the

graph to improve readability. All the articles considered were archived in the COMPASSS bibliography: http://compasss.org

7. For further discussion, see the debate between Thiem (2014) and Hug (2014). On uncertainty in QCA analysis, see Braumoeller (2015).
8. Readers will find a rigorous discussion of the conditions under which submodels preserve configurational correctness in Baumgartner and Thiem (2017b:5-9). The QCApro library for R includes a software routine to extract all correctness-preserving submodels from reference models and QCA solutions (Thiem 2018).
9. A QCA solution can be both complete and wrong simultaneously if the solution is a superset of the true model. Because some submodels are related to one another (e.g., when they are constituent components of the same complex condition), these measures could punish more severely QCA solutions which are both incorrect and complex (i.e., incorrect solutions with many submodels). Nevertheless, completeness and wrongness remain useful measures of solution fitness, built to answer a straightforward question: Among all the causal claims which are authorized by a given model, what is the proportion of claims which are actually compatible with the truth? Note that root mean squared error measure that I propose below is unaffected by the interdependence of solution submodels.
10. I use the Enhanced Quine–McCluskey Algorithm (eQMC) implemented in the eQMC function of the QCApro library for R.
11. This is analogous to the fundamental problem of causal inference that is often highlighted in the statistics literature (see Imbens and Rubin 2015). Also see Baumgartner and Thiem (2017b:9) on the configurational homogeneity condition.
12. The choice of a log-normal distribution is arbitrary. The goal is simply to offer a contrast to the constant weights assumption, which is highly unrealistic and yet is still widely adopted in simulation-based tests of QCA. Readers can easily conduct tests using alternative distributional assumptions by modifying a single line of the R code which accompanies this article.
13. On average, flipping the value of two random data points (e.g., the value of Z for one unit of observation, and the value of A for another) will produce about one truth-breaking configuration. Flipping the value of five random data points will, on average, produce about two truth-breaking configurations.
14. In theory, the maximum completeness and the minimum wrongness could come from different model candidates, but this is rarely the case in practice.
15. Again, we only consider the best available solution in any given model space.
16. At the time of writing, most of the popular implementations of crisp set QCA use 1.00 as the default consistency threshold.
17. Interestingly, with limited diversity and small sample sizes, the level of completeness can be improved by measurement error. This counterintuitive phenomenon deserves further study. One possible explanation is that the observations

introduced by mistake artificially increase the diversity of the data and produce "small" submodels that are consistent with the truth.

References

Baumgartner, Michael and Mathias Ambühl. 2019. "Causal Modeling with Multivalue and Fuzzy-set Coincidence Analysis." *Political Science Research and Methods*. doi:10.1017/psrm.2018.45

Baumgartner, Michael and Alrik Thiem. 2017a. "Model Ambiguities in Configurational Comparative Research." *Sociological Methods & Research* 46(4): 954-87. doi:10.1177/0049124115610351

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

Berg-Schlosser, Dirk, Gisèle De Meur, Benoît Rihoux, and Charles C. Ragin. 2009. "Qualitative Comparative Analysis (QCA) as an Approach." *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques* 1:18.

Braumoeller, Bear F. 2015. "Guarding Against False Positives in Qualitative Comparative Analysis." *Political Analysis* 23(4):471-87.

Huber, Peter J. 1981. *Robust Statistics*. New York: John Wiley.

Hug, Simon. 2013. "Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference." *Political Analysis* 21(2):252-65.

Hug, Simon. 2014. "We Need an Open Discussion of QCA's Limitations: A Comment on Thiem." *Qualitative and Multi-Method Research* 12(2):24-27.

Imbens, Guido W. and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, England: Cambridge University Press.

Kerner, Andrew, Morten Jerven, and Alison Beatty. 2017. "Does It Pay to Be Poor? Testing for Systematically Underreported Gni Estimates." *The Review of International Organizations* 12(1):1-38.

Krogslund, Chris, Donghyun Danny Choi, and Mathias Poertner. 2014. "Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA." *Political Analysis* 23(1):21-41.

Linsi, Lukas and Daniel K. Mügge. 2019. "Globalization and the Growing Defects of International Economic Statistics." *Review of International Political Economy* 1-23. doi:10.1080/09692290.2018.1560353

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

Mezue, Bryan. 2014. "Nigeria's Gdp Just Doubled on Paper: What It Means in Practice." *Harvard Business Review*. Retrieved October 25, 2019. (https://hbr.org/2014/04/nigerias-gdp-just-doubled-on-paper-what-it-means-in-practice).

Ragin, Charles C. 1987. *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ragin, Charles C. 2006. "Set Relations in Social Research: Evaluating Their Consistency and Coverage." *Political Analysis* 14(3):291-310.

Ragin, Charles C. 2009. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Reinhart, Carmen M. and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100(2):573-78.

Rohlfing, Ingo. 2015. "Mind the Gap: A Review of Simulation Designs for Qualitative Comparative Analysis." *Research & Politics* 2(4). doi:10.1177/2053168015623562

Roig-Tierno, Norat, Tomas F. Gonzalez-Cruz, and Jordi Llopis-Martinez. 2017. "An Overview of Qualitative Comparative Analysis: A Bibliometric Analysis." *Journal of Innovation & Knowledge* 2(1):15-23.

Schneider, Carsten Q. and Claudius Wagemann. 2012. *Set-theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge, England: Cambridge University Press.

Seawright, Jason. 2014. "Comment: Limited Diversity and the Unreliability of QCA." *Sociological Methodology* 44(1):118-21.

Simon, Herbert A. 1954. "Spurious Correlation: A Causal Interpretation." *Journal of the American Statistical Association* 49(267):467-79.

Sprites, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction and Search*. Cambridge: MIT Press.

Thiem, Alrik. 2014. "Mill's Methods, Induction and Case Sensitivity in Qualitative Comparative Analysis: A Comment on Hug (2013)." *Qualitative & Multi-Method Research* 12(2):19-24.

Thiem, Alrik. 2018. "Advanced Functionality for Performing and Evaluating Qualitative Comparative Analysis." Retrieved October 25, 2019. (http://www.alrik-thiem.net/software/).

Thomas, Herndon, Ash Michael, and Pollin Robert. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal of Economics* 38(2):257-79.

Author Biography

Vincent Arel-Bundock is an Associate Professor of Political Science. He writes on the politics of international taxation, foreign direct investment, comparative and international political economy, and research methods.