Mind the gap: A review of simulation designs for QCA - Appendix

Ingo Rohlfing*

Thursday 10th September, 2015

*Bremen International Graduate School of Social Sciences (BIGSSS), University of Bremen and Jacobs University Bremen (rohlfing@bigsss.uni-bremen.de). This work is part of the ERC Starting Grant project *Enhanced Qualitative and Multi-Method Research* (638425).

This is the appendix to the manuscript, "Mind the gap: A review of simulation designs for QCA" and elaborates on selected points of that manuscript. A general note pertains to the scope of my analysis. The analysis by Krogslund *et al.* (KCP henceforth, 2015) is broader because they assess more models of the original analyses they reproduce. I focus on the models and simulations reported in the KCP article because this suffices to illustrate my methodological points that generalize to KCP's simulations I do not directly address.

Three more specific issues that are addressed in the following are: the role of the maximum inclusion threshold (section 1), an extended discussion of the example in the manuscript and histograms summarizing the degree of model ambiguity (section 2), and, some details on my simulations of the consequences of overfitted truth tables (section 3).

1. The role of the maximum inclusion threshold

In the manuscript, I explain that it is not meaningful to assess the robustness of QCA results to variations in the *maximum* inclusion threshold. The *minimum* threshold always has to be set and distinguishes truth table rows that receive a "1" on the outcome from rows that are not assigned a "1". All rows with an inclusion score (or consistency score) that is higher than the minimum threshold receive a "1". The specification of the maximum threshold is optional and distinguishes rows that are assigned a "0" from rows that are not. All rows with a consistency value below the selected threshold are scored "0". A truth table row receives a "C", denoting 'contradiction', when both thresholds are specified and the inclusion value of a truth table row lies between both cut-off points.

One might argue that QCA researchers should set both thresholds and that rows designated as contradictions reflect uncertainty about whether they are linked to the presence or absence of the outcome. In the manuscript, I explain that the maximum threshold is of no consequence to the derivation of QCA solutions because rows valued "0" and "C" do not feed into the solution for the presence of the outcome. In addition, it holds for fuzzy sets in that in analyses of the presence of the outcome, the inclusion values of truth tables rows are *not* equal to 1 minus the inclusion value of the same row with regard to the negated outcome. The proper assignment of "0"'s to rows requires it to specify a separate truth table for the negated outcome.

If rows that are assigned as contradictions are meant to express uncertainty about whether or not they

are linked to the outcome, it would be necessary to take the uncertainty into account in the empirical analysis and derive the solution for different maximum inclusion thresholds. Whether or not the truth tables rows about which we are uncertain are valued "0" or "C" is inconsequential for the robustness analysis. In conclusion, the maximum threshold should be put aside in simulations on the robustness of QCA.

2. Model ambiguity

In the manuscript, I exemplify the meaning and mishandling of model ambiguity with one simulation on the Koenig-Archibugi data. Here, I present the complementary prime implicant chart because it highlights the presence of model ambiguity (table 1; code for deriving the two models is part of the complementary script.).[1] The chart relates the prime implicants in the rows to the truth table rows (or primitive expressions) represented by their running number in the columns. "x" denotes that a prime implicant is a superset of the configuration in the corresponding column and a "-" that the two do not stand in a set relation. The goal of the truth table analysis is to find the least complex solution, given a specific solution type, that covers all primitive expressions with at least one prime implicant.

As I explain in the manuscript, the first three prime implicants are *essential* because they cover at least one truth table row not covered by any other prime implicant. The last two prime implicants are *inessential* because only one of the two is needed to cover all primitive expressions in union with the first three primitive expressions. The fact that nonessential prime implicants figure in the analysis means that *two* models fit the information in the truth table equally well (see the complementary R script producing both models). The first model is the logical union of the first four rows in table 1, while the second model combines the first three rows and row five.

The degree of model ambiguity is in the simulations is summarized in figure 1, 2 and 3. All three simulations are described by ambiguity that is relatively small for the Ahn/Lee and Koenig-Archibugi study and extensive for Samford's QCA.

[1] Uppercase letters denote the presence of a set. Lowercase letters represent the absence of a set.

Figure 1: Number of models per simulation for Ahn/Lee (2012)

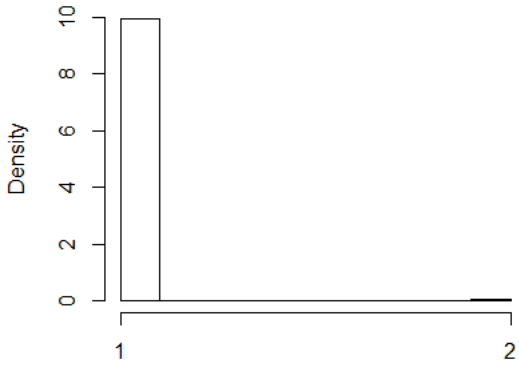

Figure 2: Number of models per simulation for Samford (2010)

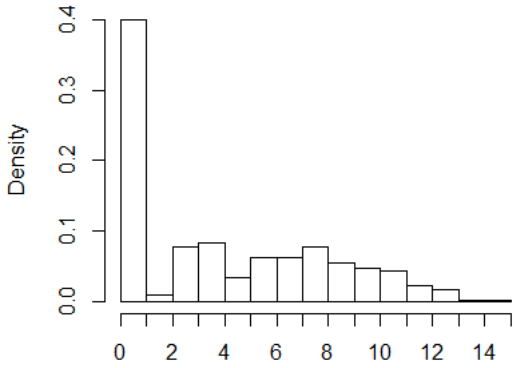

Figure 3: Number of models per simulation for Koenig-Archibugi (2004)

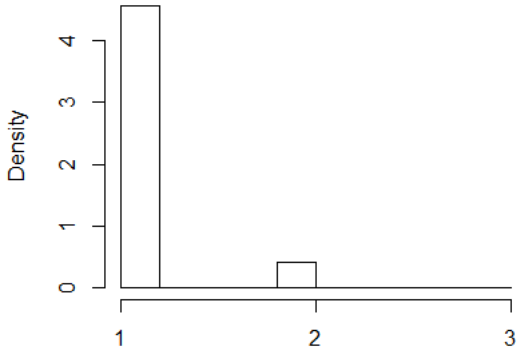

Table 1: Model ambiguity in Koenig-Archibugi simulation

Prime implicants	Primitive expressions								
	15	7	8	16	13	1	10	5	2
*CONF * capab*	x	x	-	-	x	-	-	x	-
*CONF * REGIO*	x	x	x	x	-	-	-	-	-
*conf * regio * CAPAB*	-	-	-	-	-	-	x	-	x
*idmass * regio * conf*	-	-	-	-	-	x	-	-	x
*idmass * regio * capab*	-	-	-	-	-	x	-	x	-

3. Simulations of overfitted truth tables

The three simulations are based on the reproduction data posted by KCP (2014). My reproductions of the original analyses by Koenig-Archibugi (2004), Samford (2010), and Ahn & Lee (2012) is included in the *R* script complementing my article [[[anonymized data citation]]]. Two notes are in order for the Koenig-Archibugi study and the Ahn/Lee study.

KCP choose the following target model for the Koenig-Archibugi simulation (2015, 37): $REGIO * CONF + REGIO * IDMASS * capab \Rightarrow SUPRA$. Indeed, this is the model that Koenig-Archibugi presents in his fsQCA study (2004, 162) and that I use in the main text. However, this solution is not one of the three standard solutions - conservative, intermediate, parsimonious - that one can derive with the Quine-McCluskey algorithm. Koenig-Archibugi first derives the conservative solution $REGIO * CONF \Rightarrow SUPRA$ (2004, 161). In the next step, he weighs four simplifying assumptions and decides to make only one concerning the truth table row $IDMASS * conf * REGIO * capab$. Table 2 presents the procedure by complementing the truth table with two columns. The conservative solution is based on the simplification of the first five rows. Column 'SA' denotes the simplifying assumptions that would be made if one were deriving the parsimonious solution $REGIO \Rightarrow SUPRA$.

The column 'SA made' denotes the one simplifying counterfactual Koenig-Archibugi does make (row 11). Because he makes only one simplifying assumption out of four that are required for deriving the parsimonious solution, it is *impossible* to reproduce the target solution by mechanically applying the algorithm to the original data. One can produce Koenig-Archibugi's real solution by manually adding a hypothetical case that is a member of row 11 and coding it "1" on the outcome. The conservative solution should then be derived for this truth table. In my simulation, I use the modified Koenig-Archibugi data.

Table 2: Truth table and made and non-made simplifying assumptions for Koenig-Archibugi (2004)

	IDMASS	CONF	REGIO	CAPAB	SUPRA (Y)	n	incl	SA	SA made
15	1	1	1	0	1	2	1	-	-
7	0	1	1	0	1	1	1	-	-
8	0	1	1	1	1	1	1	-	-
16	1	1	1	1	1	1	1	-	-
13	1	1	0	0	0	2	0.98	-	-
1	0	0	0	0	0	1	0.92	-	-
10	1	0	0	1	0	1	0.82	-	-
5	0	1	0	0	0	3	0.74	-	-
2	0	0	0	1	0	1	0.55	-	-
3	0	0	1	0	?	0	0.92	Y	N
4	0	0	1	1	?	0	0.91	Y	N
6	0	1	0	1	?	0	1	N	-
9	1	0	0	0	?	0	1	N	-
11	1	0	1	0	?	0	1	Y	Y
12	1	0	1	1	?	0	0.94	Y	N
14	1	1	0	1	?	0	1	N	-

For the Ahn/Lee study, KCP post a dataset comprising multiple countries in their Dataverse (2014) for extracting the data for Korea. The dataset on Korea only yields 21 observations which are fewer than Ahn/Lee draw on in their analysis. (Personal correspondence with Sophia Lee, 28.07.2015, confirmed that the available data is not the full data Ahn/Lee used). I could not access the current Korean dataset from the Comparative Welfare State Project (http://www.lisdatacenter.org/resources/other-databases/), as it is not publicly available (as of 03.09.2015) and I could not obtain it from the authors. However, the goal of my simulation is to evaluate reproduction successes when the truth table is overfitted. For this methods-centered purpose, I can use the available data for calculating reproduction successes, although this makes it impossible to specifically evaluate the robustness of Ahn/Lee's original analysis.

References

Ahn, Sang-Hoon, & Lee, Sophia Seung-yoon. 2012. Explaining Korean Welfare State Development with New Empirical Data and Methods. *Asian Social Work and Policy Review*, **6**(2), 67–85.

Koenig-Archibugi, Mathias. 2004. Explaining Government Preferences for Institutional Change in EU Foreign and Security Policy. *International Organization*, **58**(1), 137–174.

Krogslund, Chris, Choi, Donghyun Danny, & Poertner, Mathias. 2014. Replication data

for: Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA. *http://dx.doi.org/10.7910/DVN/27100*.

Krogslund, Chris, Choi, Donghyun Danny, & Poertner, Mathias. 2015. Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA. *Political Analysis*, **23**(1), 21–41. Published.

Samford, Steven. 2010. Averting "Disruption and Reversal": Reassessing the Logic of Rapid Trade Reform in Latin America. *Politics & Society*, **38**(3), 373–407.