

## „Using Multi-Value Logic Synthesis in Social Science“

Paper prepared for the 2<sup>nd</sup> General Conference of the European Consortium for Political Research (ECPR), Marburg, September 2003. Section 6: *Methodological Advances in Comparative Research : Concepts, Techniques, Applications*, Panel 8: *Assessing the respective potential of Qualitative Comparative Analysis (QCA), Fuzzy Sets and other techniques : applications.*

Lasse Cronqvist  
Institute of Political Science  
Philipps-University,  
Marburg/Germany

E-Mail: [lasse@staff.uni-marburg.de](mailto:lasse@staff.uni-marburg.de)

### Abstract:

One of the main problems using Qualitative Comparative Analysis has been the restriction to use only dichotomised variables. In this paper an extension to QCA is presented, which allows the use of multi-valued variables with Boolean synthesis, and the problem of setting the thresholds is discussed to overcome this methodological problem. A new software named TOSMANA (Tool for Small-N Analysis) is introduced which allows the use of Multi-Value QCA (MVQCA). Also some of the problems of setting threshold used with QCA and MVQCA are discussed.

## 1. Introduction<sup>1</sup>

QCA has been used for research on several topics of political science and abroad (see DeMeur / Rihoux: (2002) for a listing of some examples), but the method also has its shortcomings. One of the main problems with QCA has been stated as the compulsory transformation of data to dichotomy data (Berg-Schlosser 2002), which bears the risk of data loss and may create a large number of contradicting case configurations, which diminish the use of the analysis.

In FS/QCA Ragin and Drass (2002) implemented the fuzzy-set approach presented by Ragin (2000) to overcome the dichotomy problem. With fuzzy-sets it is possible to assign values between 0 (fully out) and 1 (fully in) to each variable to a case, giving the possibility to create a more detailed case configuration. These configurations are then processed with proportional methods, to find the necessary and sufficient conditions to explain an outcome.

In this paper, an other approach for the use of non-dichotomy data in social science is introduced: The Multi-Value Qualitative Comparative Analysis (MVQCA) keeps the idea of minimizing expressions with Boolean Analysis by allowing non-Boolean literals to be included in the analysis. This means that a truth function is assigned for each literal, so that the literals can be combined like Boolean literals in QCA are being combined. This allows the use of Multi-value scales within one variable, and then to use them to minimize complex data sets to parsimonious expressions. MVQCA thereby allows the use of Multi-value variables as they are also used in many fuzzy-set applications (for example Ragin 2000: 271-307, Pennings 2003), but instead of apply an analysis using probabilistic criteria to identify sufficient and necessary conditions, the veristic synthesis of data sets used in QCA is still used with MVQCA.

MVQCA uses the ideas of Multi-value Logic Synthesis developed by the electronic engineers at Berkeley University (see Brayton / Sunil 1999), which have used Multi-value synthesis for computer aided circuit design. Their papers on this topic are a good help to access the mathematical basics of the topics. Although there is also a software program provided by the researchers at the Berkeley University (see MVSIS Group), a new software called TOSMANA (Tool for Small-N Analysis) has been developed for the use of MVQCA (Cronqvist 2003a).

---

<sup>1</sup> Acknowledgement: I wish to thank Prof. Dr. Dirk Berg-Schlosser for his many helpful comments on the draft of this paper.

This is mainly due to guarantee that further developments on (MV)QCA as well as other methods of macro-qualitative social science can be implemented easily, and because specific topics related to social science data analysis do not exist in electronic engineering applications (like ordinal variables just to mention one).

### *This paper*

In his paper on the shortages of the case-oriented approach as well as of the variable-oriented approach, Goldthorpe (1997) explains that the limitation of QCA to dichotomous coding of data is a further restriction to social research. The high sensitivity of calculation toward the threshold setting inflicts the power of the analysis in a severe way (Goldthorpe 1997:3), and, to use the “QCA-language”, a slightly change of the threshold may result in a different solution of the analysis. In fact, using “unlucky” thresholds can place the scientist on a dead end road, if the result of the data analysis is used without further analysis.

In this paper I want to introduce an extension of QCA with multi value variables. I will start with a short introduction of the subject of the subject of Multi-Value Qualitative Comparative Analysis (MVQCA). Without going into deep technical details, the basic features of this method will be explained, and these will be illustrated by some examples. Secondly, the potential of MVQCA to diminish the consequences of re-coding interval scaled variables to ordinal scales with only a few scale values, will be shown by adapting a data set earlier used for QCA analysis to MVQCA methods, which can reduce the number of contradictions considerably. Finally, I will make some remarks on "threshold setting" - the question how thresholds can be set in a responsibly way. All this thoughts are also present in the new software for (MV)QCA - TOSMANA. More information on how to use the TOSMANA software can be found in the TOSMANA User Manual (Cronqvist 2003b).

## **Multi Value Qualitative Comparative Analysis (MVQCA)**

First, the main features of the MVQCA method will be described. Readers familiar with the “Qualitative Comparative Analysis” developed for the social sciences by Charles Ragin (see Ragin 1987) will discover some similarities, but also some fundamental differences between these methods.

### ***Basic Features of MVQCA***

Multi-Value analysis requires that each variable is available on either an ordinal or a nominal basis. As an example, the variable “region”, which contains information about the geographical location of a country, is a nominally scaled variable, while the variable “GDP-Growth” with the possible values of  $P_{GDP} = \{\text{negative, low or zero, medium, high, very high}\}$  is an example of a five-step ordinal scale. This means that raw interval scaled data can not be used for MVQCA directly, but must be transformed into values inside a multi value set. This is done by selecting an array of thresholds according to which the values are assigned. This, of course, implies a certain loss of more finer-graded information. This means that there may be no differences between two cases in the transformed data, although there might be a (more or less small) difference in the original raw data. Therefore the choice of thresholds has to be made very carefully, a subject to be discussed in depth further below.

### ***Multi-Value Notation***

As multi-value variables can have more than two states, set notation has to be used to represent the cases and implicants as formal expressions. Each expressions consists of one or more literals  $X_i^{S_n}$  where  $X_i$  is a variable from the data set and  $S_n$  is a set of values of the variable valid in this expression.  $I=A\{0,1\}$  indicates that the implicant  $I$  represents all cases having a value of  $A$  which is either  $0$  or  $1$ . To make the reading more comfortable, the notation  $I=A_{0,1}$  is used in this text as well. The notation  $I=A_{0,2}$  may be used to indicate  $A$  having all values between  $0$  and  $2$  ( $= 0,1,2$ ), and  $I=A_{\sim 2}$  means all values except for  $2$ .<sup>2</sup>

---

<sup>2</sup> Researchers used to Boolean QCA will see that this notation can also be used to express Boolean terms: Instead of using upper case letters ( $A$ ) for indicating the presence of a condition and lower case letters ( $a$ ) for the absence of a condition, the value  $0$  ( $A_0$ ) will be used for indicating the absence of a condition and the value  $1$  ( $A_1$ ) for the presence of a condition.

*Combining literals and implicants.*

Very rarely a data set can be explained by just one literal. Almost always explanations have to be found in combined implicants, consisting of more than one literal, or even multiple, combined implicants. For this Boolean algebra is used. Each literal  $X_i^{S_n}$  represents a number of cases, which have one of the states in the variable  $X_i$  indicated by  $S_n$ . For those cases  $C_i$  the Boolean Truth function is  $\pi_{X_i^{S_n}}(C_i) = T$  (true), whilst  $\pi$  is F (false) for other cases. The following simple example will illustrate this:

Case Number	Value of A	Value of B	Value of C	$\pi_A^{(0)}(C)$
1	0	1	1	T
2	1	2	1	F
3	2	2	1	F
4	0	1	2	T

**Table 1: Example 1. Multi Value Cases and the Boolean Truth function**

The Boolean Truth function of the literal  $A_0$  is True for the cases 1 and 4 as the value of A those cases is exactly 0, while the function is False in the cases 2 and 3 as the value of A is not 0.

*Boolean AND*

To combine multiple literals into one expression, the Boolean *AND* is used. This is sometimes called Boolean multiplication (Ragin 1987: 91), which may be confusing as it is substantially different from arithmetic multiplication. If an expression consists of two literals combined by the Boolean *and* (the sign  $\otimes$  is used here), the expression is only true if the Boolean Truth function of both literals with the specific case is true. This means that in the upper example the Boolean Truth function of the expression  $A_0 \otimes C_1$  is only true for case 1, as only case 1 is having the value 0 for variable A and 1 for the variable C. In standard use of MVQCA, the sign  $\otimes$  is mostly left out, and the expression  $A_0 \otimes C_1$  is written as  $A_0 C_1$ .

$A \otimes B$	T	F
T	T	F
F	F	F

Figure 1. Boolean *AND*

$A \oplus B$	T	F
T	T	T
F	T	F

Figure 2. Boolean *OR*

### *Boolean OR*

To combine two or more Boolean expressions, the Boolean *OR* is used, which sometimes also is called Boolean addition. If a Boolean expression consists of two Boolean expressions combined by *OR* ( $\oplus$ ), the Boolean Truth function of the combined expression is true if one or both Boolean Truth functions of the expressions used to combine the new expression is true. In example 1 the Boolean Truth function of the combined expression  $A_0B_1 \oplus A_1B_2$  is True for the cases 1, 2, and 4 as the first expression of the addition is true for case 1 and 4 and the latter for case 2.

In daily use, the *OR* sign  $\oplus$  can be replaced by the normal addition + sign, if it is clear that the Boolean addition is used.

### *Representing Data*

To use Multi-Value coded data it is necessary to construct the data as a data matrix. Unlike the truth tables used by QCA, the data variables used are not restricted to binary, nominal-scale variables (Ragin 1987: 87), but can be non-binary and of an ordinal type, too. Also, not all possible logical combinations are represented in the matrix. In fact this would very quickly exhaust the available resources, as we would run into a very high number of possible combinations. Considering a simple example should make this clear:

As an example for Qualitative Comparative Analysis Lipset's social indicators explaining the democratic breakdown in Interwar Europe have been used (Berg-Schlosser and De Meur 1994: 255-257). Lipset employs four socio-economic indicators "sustaining" democracy. These are: Gross National Product (*G*), Degree of Urbanisation (*U*), Literacy (*L*), and Industrial Labour Force (*I*). If the variables are transformed into Boolean variables, there are 16 ( $2*2*2*2$ ) logical combinations. In fact, the number of logical combinations in a truth table is given by the number of variables, and calculated with the formula  $\eta=2^{|v|}$  where  $|v|$  indicates the number of variables. In MVQCA the number of combinations is higher for the same number of variables, as every variable may have more than two values assigned.<sup>3</sup>

---

<sup>3</sup> Please note that the range of values can be selected for each variable independently, and that they may differ from each other. For MVQCA, the numbers  $\{0, \dots, n-1\}$  have to be used to construct the range of the size  $n$ .

<i>Variable name</i>	<i>Possible values (P)</i>	<i>Number of Variables (v<sub>i</sub>)</i>
<i>GDP</i>	<i>very low, low, medium, high, very high</i>	<i>5</i>
<i>Urbanisation</i>	<i>very low, low, medium, high, very high</i>	<i>5</i>
<i>Literacy</i>	<i>low, medium, high, almost 100%</i>	<i>4</i>
<i>Industrialisation</i>	<i>low, medium, high</i>	<i>3</i>

**Table 2: Number of Variables in Multi-Value Sets**

In table 1.1. the Lipset indicators are assigned (exemplary) Multi-Value ranges. In this example there are 300 ( $5*5*4*3$ ) possible combinations to deal with. The number of combinations is given by the formula  $\eta=|v_1| \times \dots \times |v_n|$ , meaning that you have to multiply the number of possible values of each variable to get the number of possible combinations. In fact it is possible to recode multi value sets to binary sets using dummy variables<sup>4</sup> but given the very high number of possibilities resulting from this recoding, this approach seems to be less useable.<sup>5</sup> Also, the limitation to a rather small number of variables in the common software used for QCA does not allow the usage of middle size MVQCA data sets.<sup>6</sup> Therefore a new algorithm has been included in the TOSMANA Software Package (Cronqvist 2003a), which is using a multi-value set algorithm to calculate the minimized expressions (see Cronqvist 2002).

### *Parsimony*

The aim of MVQCA is (similar to QCA) to find *parsimonious explanations* (Ragin 1987: 83) in complex data sets, and thereby to specify variables of major importance. Unlike

<sup>4</sup> This possibility is given by the Espresso algorithm included in FS/QCA (Ragin/Drass 2002).

<sup>5</sup> In fact, there are a number of problems by encoding multi value data to binary data (Brayton and Khatri 1999: 3): The main disadvantage is the number of additional variables added, by which the number of don't cares rises to very high levels. This is mainly because a lot of combinations are created which will never occur. Imagine a traffic light: To make this example easy, the traffic light has three possible values: Red light, yellow light, and green light. Encoding the Multi-Value variable *Light* with  $P_{Light}=\{red,yellow,green\}$  into three binary dummy variables (red,yellow,green) would result in the following table:

<u>MV Value</u>	<u>red value</u>	<u>yellow value</u>	<u>green value</u>
Red	1	0	0
Yellow	0	1	0
Green	0	0	1

Now the point is that not only the upper (real) combinations will be created, but also all other possibilities will be constructed. In reality, a combination like red=1 and green=1 will never occur in regular traffic light usage, but the combination still occurs in the data matrix when including remainder cases in the analysis. This means that a high number of never used combinations will be created and combined with the other independent variables, and these must all be included in the analysis.

<sup>6</sup> In QCA3.0 (Drass/Ragin 1992) the number of variables was restricted to 12. In FS/QCA (Ragin/Drass) there is no restriction to the number of variables included, but the software produces an error if more than 25 variables are used. Of course this is no harm when dichotomious variables are used, as this number of variables would not be used often. But when using dummy variables for multi-value reduction, the restriction to 25 dummy variables would mean a restriction to a smaller number of variables, as every multi value variable has to be recoded to multiple binary variables.

quantitative, statistical methods, it is possible to discover new connections between variables instead of only checking “old” connections. (Ragin 1987: 84). The main principles of MVQCA will be described by two imaginary data sets:

<i>Variable name</i>	<i>Possible values (P)</i>	<i>Number of Variables</i>
<i>A</i>	<i>0,1</i>	<i>2</i>
<i>B</i>	<i>0,1,2</i>	<i>3</i>
<i>C</i>	<i>0,1,2</i>	<i>3</i>
<i>Outcome</i>	<i>0,1,2</i>	<i>3</i>

**Table 3. Variables in example data set (Example 2+3).**

<i>Case</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>Outcome</i>
<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>2</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>0</i>
<i>3</i>	<i>0</i>	<i>0</i>	<i>2</i>	<i>0</i>
<i>4</i>	<i>0</i>	<i>1</i>	<i>0</i>	<i>0</i>
<i>5</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>
<i>6</i>	<i>0</i>	<i>1</i>	<i>2</i>	<i>1</i>
<i>7</i>	<i>0</i>	<i>2</i>	<i>0</i>	<i>2</i>
<i>8</i>	<i>0</i>	<i>2</i>	<i>1</i>	<i>1</i>
<i>9</i>	<i>0</i>	<i>2</i>	<i>2</i>	<i>2</i>
<i>10</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>1</i>
<i>11</i>	<i>1</i>	<i>0</i>	<i>1</i>	<i>0</i>
<i>12</i>	<i>1</i>	<i>0</i>	<i>2</i>	<i>1</i>
<i>13</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>1</i>
<i>14</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>
<i>15</i>	<i>1</i>	<i>1</i>	<i>2</i>	<i>2</i>
<i>16</i>	<i>1</i>	<i>2</i>	<i>0</i>	<i>1</i>
<i>17</i>	<i>1</i>	<i>2</i>	<i>1</i>	<i>2</i>
<i>18</i>	<i>1</i>	<i>2</i>	<i>2</i>	<i>2</i>

**Table 4. Example data set (Example 2) .**

### *Minimization*

The goal of MVQCA is to extract short (parsimonious) explanations of an outcome by minimizing the complex multi-value data set. In QCA, minimization is rather easy: If two expressions only differ in one term, these two can be taken to produce a combined, reduced expression. If two logical combinations both have the outcome 0 and their expressions are given by *abc* and *Abc*, then *A* is irrelevant, and the combined expression can be stated as *bc*. As indicated above, multi-value synthesis is a generalisation of Boolean synthesis: A number of expressions can only be replaced by a reduced expression if all expressions in the data including the reduced expression have the same outcome.<sup>7</sup> In example 2 the expression  $A_0B_0C_0$ ,  $A_0B_0C_1$ , and  $A_0B_0C_2$  can be reduced to  $A_0B_0$  because all three expressions in the data sets implied by this reduced expression have the same outcome 0.

<sup>7</sup> If an expression  $\epsilon'$  is included in  $\epsilon$ , this means that all variable values included in  $\epsilon'$  are also included in  $\epsilon$ . Then it can also be said that  $\epsilon'$  implies  $\epsilon$ .

In other words, the most fundamental rule of Boolean reduction as expressed by Ragin can be rewritten for multi-value reduction:

(1) “If two Boolean expressions differ in only one causal condition yet produce the same outcome, then the causal condition that distinguishes the two expressions can be considered irrelevant and can be removed to create a simpler, combined expression.” (Ragin 1987: 93).

For multi-value reduction this can be written as:<sup>8</sup>

(2) If all  $n$  multi-value expressions ( $c_0\Phi, \dots, c_{n-1}\Phi$ ) differ only in the causal condition  $C$  while all  $n$  possible values of  $c$  yet produce the same outcome, then the causal condition  $C$  that distinguishes these  $n$  expressions can be considered irrelevant and can be removed to create a simpler, combined expression  $\Phi$ .

The reader should convince himself that the rule for Boolean reduction is a specialisation of the rule for multi-value reduction, and that the rule for multi-value reduction also is valid for Boolean reduction.

But as the number of don't cares will be very high in data sets with many fine-graded scales, minimizing cases with the same outcome without including don't cares would be a very short event, if such a data set is used. To remove a causal condition requires that *all* expressions implied by this condition produce the same outcome. In the Lipset data set mentioned above, this would mean that to remove  $G$ , all expressions with  $G$  implied by the new specified expression  $U_xL_yI_z$  would have to have the same output. This might be possible for one single combined expression, but as minimizing should move beyond the point of single reductions and we want to reduce already combined expressions, this strategy seems non practicable: You would need at least 25 (5 possible values for  $G$  \* 5 possible values for  $U$ ) cases to be able to get a combined expression of  $L_yI_z$  with the same outcome. Considering the low number of cases in small-N Analysis this shows the problem clearly. Therefore, cases that exist logically but not in the data set - the “remainder cases” – will be included in the synthesis. This creates a new rule for multi-value reduction including remainder cases:

(3) If all  $n$  multi-value expressions ( $c_0\Phi, \dots, c_{n-1}\Phi$ ) differ only in the causal condition  $C$  with  $n$  possible values yet produce the same outcome or are non existing expression (remainders), then the causal condition  $C$  that distinguishes these  $n$  expressions can be considered irrelevant and can be removed to create a simpler, combined expression  $\Phi$ .

---

<sup>8</sup>  $\Phi$  indicates an combined expression.

But as there are only three groups of expressions (Cases with outcome  $o$ , Cases with an outcome other than  $o$ , remainders)<sup>9</sup>, this means the rule can be changed to:

(4) If two or more multi-value expressions  $c_i\Phi \in \{c_0\Phi, \dots, c_{n-1}\Phi\}$  differ only in the causal condition  $C$  with  $n$  possible values yet produce the same outcome, then the causal condition  $C$  that distinguishes these  $n$  expressions can be considered irrelevant and can be removed to create a simpler, combined expression, if there is no expression implied by the new expression  $\Phi$  producing a different outcome.

In fact, even if only one expression is implied by the expression  $\Phi$  multiplied with  $C$ ,  $C$  can be removed if no expression implied by  $\Phi$  is producing a different outcome.

<i>Case</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>Outcome</i>
1	0	0	0	0
2	0	0	1	0
3	0	0	2	0
4	0	1	0	0
5	0	2	2	2
6	1	0	0	1
7	1	0	1	0
8	1	2	0	1
9	1	2	1	2

**Table 5 Example data set (Example 3)**

In Example 3, there is a number of combinations with no outcome assigned (don't cares), which are not stated. For the outcome 0 using the upper rule (4), we find that the cases 1-3 can be reduced to  $A_0B_0$ . Also, case 7 ( $A_1B_0C_1$ ) can be reduced to  $B_0C_1$  as there is no case with an other outcome implied by this expression. For the outcome 0, the following reductions are possible:

- $A_0B_0C_0, A_0B_0C_1, A_0B_0C_2$  can be reduced to  $A_0B_0$
- $A_0B_0C_0, A_0B_1C_0$  can be reduced to  $A_0C_0$
- $A_0B_0C_1, A_1B_0C_1$  can be reduced to  $B_0C_1$
- $A_0B_0C_1$  can be reduced to  $A_0C_1$
- $A_0B_0C_2$  can be reduced to  $A_0C_2$
- $A_0B_1C_0$  can be reduced to  $A_0B_1$
- $A_0B_1C_0$  can be reduced to  $B_1C_0$

It is even possible to reduce  $A_0B_1$  and  $B_1C_0$  to  $B_1$ .

<sup>9</sup> Readers with knowledge of QCA may object that there are also combinations with contradictory (different) outcomes. This is, of course, also true in MVQCA, but here such outcomes are given a special outcome value and are viewed as an independent group of cases, which may be merged with any other group or ignored, if desired.

### *Excluding Remainders*

When working with data sets only containing few non dichotomous variables, it also makes sense to exclude remainders from the reduction sometimes. Then an adopted Quine algorithm, which is following the rule set above as (2) is performed over the data set. In Example 3 the cases 1,2, and 3 can be merged, as the three cases only differ in the Variable  $C$  and they all have the same outcome. Also, all possible cases implied by the new expression resulting from the reduction are existing (and having the outcome 0). Hence it can be noted that

$$A_0B_0C_0 + A_0B_0C_1 + A_0B_0C_2 \Rightarrow A_0B_0$$

As stated, this method will show useable results with data sets with only a few multi value variables only. If too many multi valued variables are used, there will be no or only a very limited number of reduction steps possible.

### *Implicants, Prime Implicant, Cover and Prime Cover*

The idea of implicants and covers has already been used by Charles Ragin for the construction of QCA (Ragin 1987). The same concept is used with MVQCA. The goal is to find a minimal prime cover which covers all the cases with an outcome  $o$  without covering any case with an other outcome than  $o$ . To understand this, first, some definitions have to be made.<sup>10</sup>

Map: A case  $C$  is mapped by an expression  $\alpha$ , if  $\alpha$  is a subset of the expression of  $C$  ( $\alpha$  implies the expression of  $C$ ).

Set of Representation: The set of representation of an expression  $\alpha$  consists of all cases mapped by  $\alpha$ .

Implicant: An implicant is a multi-value expression, with all cases from the data set mapped in it have the outcome chosen. [= no case mapped in the implicant may have another outcome]

Prime Implicant: An expression  $\alpha$  is a prime implicant if there is no other implicant  $\alpha'$  which implies  $\alpha$ .

Cover: A cover is a set of implicants whose union of case memberships contains all the cases with the selected outcome.

---

<sup>10</sup> A mathematical definition of these definitions is given by Brayton and Khatri (1999). Readers with some in-depth knowledge of (formal) mathematics should take a look at the formal definition of multi value logic synthesis.

Prime Cover: A prime cover is a cover whose elements are all prime.

Minimal Prime Cover: A prime cover is minimal if no prime implicant can be removed from the prime cover without breaking the cover quality of the set of prime implicants.

The goal of minimizing complex multi-value data sets is achieved when a prime cover is found. To find such a prime cover, two steps of calculations are necessary: The first step has to find the prime implicants, whilst the second step has to combine these prime implicants to covers as small as possible.<sup>11</sup> As the method used in QCA, which reduces the data set from top to bottom, would imply a huge amount of memory to be used to process the calculation when using multi-value sets, another algorithmic approach has been chosen to find the prime implicants. For example, consider again the Lipset indicators (see table 2). This data set has a complexity of  $5*5*4*3 = 300$  possible configurations in the data matrix. With modern computer technology processing this may not be a problem. But only few data matrices consist of only four independent variables, often ten or more variables have to be taken into account when using MVQCA. A data set with ten variables with two variables with four values, five variable with three values and three dichotomized variables gives a mass of  $4^2*3^5*2^3 = 31104$  possible configurations. Then nearly all possible configurations have to be reduced, if the remainders are included in the analysis. So if we have a data set consisting of four variables like the Lipset indicators, and we are looking at eight cases with the outcome 0 and ten cases with the outcome 1, we have to reduce 292 cases having the outcome 1 or being don't cares to find the prime implicants for the outcome 1. If we are having the same number of cases for a ten variables data set, we are looking on 31096 cases to be reduced.<sup>12</sup> To reduce multi-value expressions, therefore, a new algorithm has to be used, which finds the prime implicants from bottom to top. Briefly stated, the calculation is done starting with the shortest possible expressions (consisting of just one literal) and extending them until they only map cases with the selected outcome and no other cases<sup>13</sup>.

Before the solutions are found next, some calculations are done to reduce the number of prime implicants. Let A and B be two nominal scaled variables. If  $A_0B_0$  and  $A_2B_0$  are both prime

---

<sup>11</sup> This is the same procedure as in Boolean QCA (Ragin 1987: 95ff)

<sup>12</sup> Also, it is not possible to use De Morgan's law to reduce multi-value expressions, as this law only is applicable to Boolean data sets.

<sup>13</sup> A Powerpoint presentation on this algorithm can be found on the author's web side (Cronqvist 2002).

implicants, then these are joined to the new prime implicant  $A_{0,2}B_0$ . This could also be done with ordinal variables, but it makes little sense to combine the two prime implicants if  $A_1B_0$  implies cases with an other outcome, so in TOSMANA values of ordinal variables are only joined, if they can be “connected” by ‘-’. In this case  $A_0B_0$  and  $A_2B_0$  will only be joined if  $A_{0,2}B_0$  is a prime implicant. This means that  $A_1B_0$  must be a prime implicant its self or that there is no case implied by  $A_1B_0$ . This is the point mentioned in the introduction where ordinal and nominal variables are treated differently.

When the prime implicants have been found and eventually reduced, these have to be combined to the shortest possible covers. TOSMANA is able to find all shortest prime covers of the smallest size  $s$  and can even be configured to find all prime covers, or just prime covers up to a certain size. In MVQCA (as well as in QCA) the size is calculated by adding the size of the prime implicants, which are calculated by the number of literals included. This means that the size of the implicant  $A_0B_0C_1$  is the same as the size of the implicant  $A_0B_0C_{1,2}$  although the latter implicant includes two values of the variable  $C$ . This way of calculating the size of a prime cover has to be discussed, but the software (TOSMANA) is constructed to adapt other modes of measuring the size of prime implicants and prime covers.

## 2. Improving QCA Analysis with Multi-Valued Scales

To indicate the use of MVQCA, I would like to show an example where changing the variable scale gives a more satisfactory result.

### ***Vanhanens Indices on Democratisation***

One of the first applications of QCA in social sciences was published by Dirk Berg-Schlosser and Gisèle Demeur (1994). The authors revisited major hypotheses on the conditions of democracy in interwar Europe and made them subject to a Boolean analysis. One of the hypotheses tested were the Indices of Democratisation by Vanhanen.

In his work on the emergence of democracy in 119 states, Vanhanen (1984) defines three variables, which "indicates three different dimensions of politically relevant power resources" which "together [...] measure the social conditions and structures on which the nature of political systems depend" (Vanhanen 1984: 36). The three variables are:

Index of Occupational Diversification (IOD): This index contains the percentage of urban population and the percentage of non-agricultural population. This variable is "assumed to indicate the distribution of economic and occupational interests and, indirectly, the distribution of economic and organizational power resources" (Vanhanen 1984: 36)

Index of Knowledge Distribution (IKD): This index contains percentage values of the number of students and literates. This index should "indicate relative differences between the countries in the distribution of intellectual power resources" (Vanhanen 1984: 36).

Share of Family Farms (FF): This value indicates the share of family farms of the total area of holdings. (Vanhanen 1984: 24).

Berg-Schlosser and De Meur (1994: 257f) examine the explanatory potential of these three indicators to explain the differences in the survival of democracy in 17 states of Interwar Europe<sup>14</sup>. Table 6 shows the original data given by Vanhanen and the dichotomised data by Berg-Schlosser / De Meur. For each variable a threshold of 50% is used, as these levels were suggested by Vanhanen (Berg-Schlosser / DeMeur 1994: 258).

---

<sup>14</sup> These cases are (with the abbreviations used further on): Austria (AU), Belgium (BE), Czechoslovakia (CS), Finland (FI), France (FR), Germany (GE), Greece (GR), Hungary (HU), Italy (IT), the Netherlands (NL), Poland (PL), Portugal (PO), Romania (RO), Spain (SP), Sweden (SW), Great Britain (GB). Although Vanhanen provides data on ten more European states the authors do not use them, as they argue that the major cases are included in the analysis by using the selected cases (Berg-Schlosser / De Meur 1994: 254).

<i>Case</i>	<i>IOD</i>	<i>D(IOD)</i>	<i>IKD</i>	<i>D (IKD)</i>	<i>FF</i>	<i>D (FF)</i>	<i>Outcome</i>
<i>AU</i>	51,5	1	55	1	45	0	0
<i>BE</i>	64	1	51,5	1	30	0	1
<i>CS</i>	38,5	0	49	0	40	0	1
<i>FI</i>	21,5	0	46,5	0	47	0	1
<i>FR</i>	48	0	50,5	1	35	0	1
<i>GE</i>	53	1	54	1	54	1	0
<i>GR</i>	34	0	28	0	28	0*	0
<i>HU</i>	37	0	47	0	40	0	0
<i>IT</i>	38	0	39,5	0	22	0	0
<i>NL</i>	61	1	51,5	1	40	0	1
<i>PL</i>	17,5	0	37,5	0	53	1	0
<i>PO</i>	30,5	0	18,5	0	20	0	0
<i>RO</i>	16,5	0	25	0	41	0	0
<i>SP</i>	35	0	33	0	20	0	0
<i>SW</i>	39,5	0	52,5	1	50	1	1
<i>GB</i>	78,5	1	50	1	25	0	1

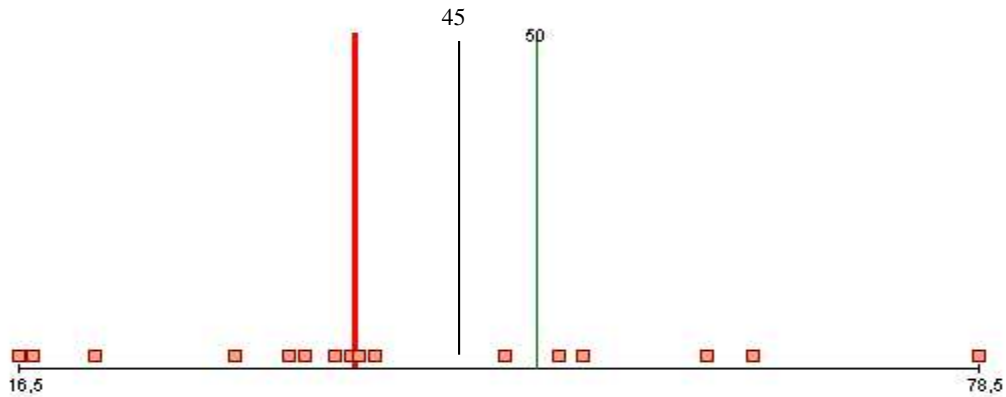
**Table 6: The data represents the values of the decade 1920-1929 for each variable in Vanhanen 1984. The dichotomised data is from Berg-Schlösser / De Meur 1984. \* Although Greece has a value of 28% of Family Farms, the dichotomised value of 1 has been used for the Boolean analysis in the analysis by Berg-Schlösser and DeMeur. This error does actually not effect the results of the analysis, as Greece would join the group of contradictory cases with all values = 0 and Poland would still be form a own group so that the QCA settings would not change. In this paper I will use the correct data.**

The decision to use a 50% threshold for all three variables was an unfortunate one, as we will see. When performing the Boolean synthesis, first we find that 12 of the 16 cases are contradicting so that only two cases with the outcome 0 and two cases with the outcome 1 are explained.

<i>Cases</i>	<i>IOD</i>	<i>IKD</i>	<i>FF</i>	<i>Outcome (QCA)</i>
<i>RO, FI, PO, GR, SP, HU, IT, CS</i>	0	0	0	<i>C</i>
<i>PL</i>	0	0	1	0
<i>SW</i>	0	1	1	1
<i>FR</i>	0	1	0	1
<i>AU, BE, NL, GB</i>	1	1	0	<i>C</i>
<i>GE</i>	1	1	1	0

**Table 7: Truthtable of the Vanhanen Indices using dichotomous data.**

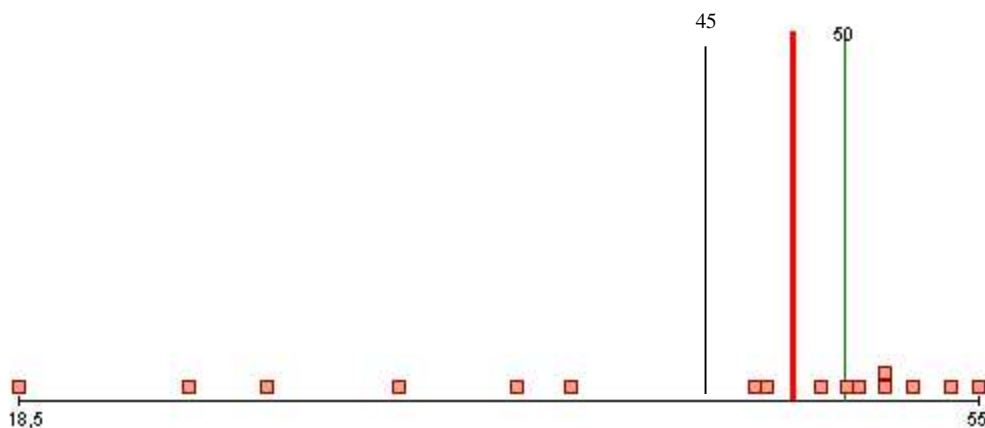
This is indeed not satisfying, and sets up a big question sign if these indices can be used to explain democracy breakdown in the interwar period in Europe. But if we look at the distribution of cases and the thresholds, we see several irregularities. As an example the variable IOD:



**Figure 3: Distribution of values of IOD. The thick line indicates the median, the narrow lines the new selected threshold of 45% and the previous threshold of 50%.**

Setting the threshold at 50% looks very randomly, as it separates two values close to each other (48%, 51,5%) and it also cuts the set of cases into the subsets of different size. A simple average linkage cluster moves the threshold to somewhere between the cases with the values of 39,5% and 48% (setting the threshold to 45% satisfies this).<sup>15</sup>

Also the threshold of 50% for IKD seems unfortunate, as it cuts right through a group of cases close to each other:



**Figure 4: Distribution of values of IKD. The thick line indicates the median, the narrow line the original selected threshold of 50% and the new threshold at 45%.**

Setting the threshold to 45% makes the division look more reasonable.

Finally the threshold of FF is difficult to place. Using a threshold of 50% cuts the set into two unbalanced subsets (containing 4 and 12 cases), but due to the distribution all along the range between 20% and 55%, it looks difficult to place the threshold without a more serious loss of

<sup>15</sup> Setting the thresholds using a cluster may also seem at least questionable, as no theoretical assumptions about the variable are included in the decision. But thresholds have to be set so that they clarify as much as possible and not create artificial groups of cases. See the chapter at the end on thresholds for more on this.

data precision. To solve this problem, a set of two thresholds are used. They are placed at 32 % and 43 % (using simple average clustering), which cuts the data set into three equal size portions: One subset with low values of FF (<32%, 5 cases), one subset with middle values (>32% and <43%, 5 cases) and one subset with high values (>43%, 6 cases).

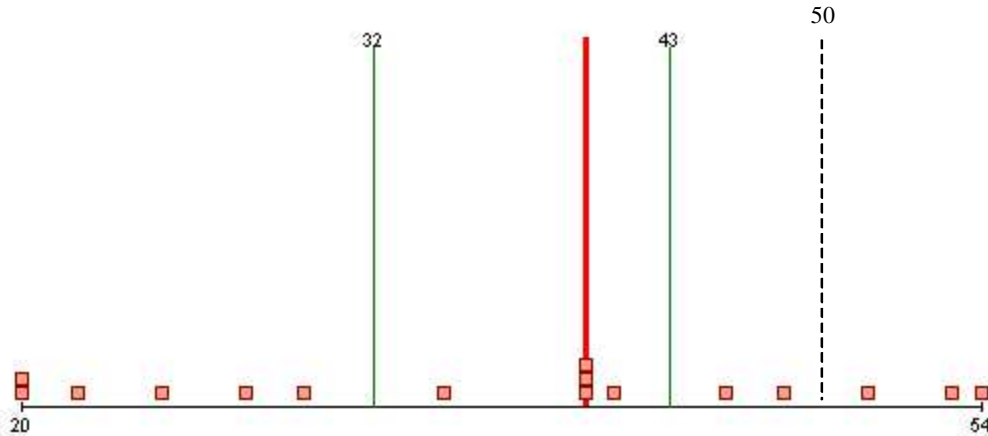


Figure 5: Distribution of values of FF. The thick line indicates the median, the narrow line the two thresholds at 32% and 43%. The dotted line indicates the original used threshold of 50% as used by Berg-Schlusser and De Meur.

Using these thresholds, only two cases contradict (CS and HU), whilst the other 14 cases are not contradicting with cases of an other outcome.

<i>Cases</i>	<i>IOD</i>	<i>IKD</i>	<i>FF</i>	<i>Outcome (QCA)</i>
<i>PO, GR, SP, IT</i>	0	0	0	0
<i>RO</i>	0	0	1	0
<i>PL</i>	0	0	2	0
<i>HU, CS</i>	0	1	1	C
<i>FI, SW</i>	0	1	2	1
<i>BE, GB</i>	1	1	0	1
<i>FR, NL</i>	1	1	1	1
<i>AU, GE</i>	1	1	2	0

Table 8: Condensed data table with the use of a multi value coding of FF and revised thresholds of IOD and IKD.

If we take a closer look on the raw data of the two cases contradicting, we see that they indeed are very similar, so that dividing the cases would only be possible with a very artificial threshold<sup>16</sup>:

<i>Case</i>	<i>IOD</i>	<i>IKD</i>	<i>FF</i>	<i>Outcome</i>
<i>HU</i>	37	47	40	0
<i>CS</i>	38.5	49	40	1

Table 9: Raw data for the two cases HU and CS still contracting with the new thresholds.

<sup>16</sup> We would have to define a threshold which cuts exactly between 37% and 38,5% on the IOD scale, not a very reasonable setting with the data used here (see chapter on how to set thresholds below).

By changing the thresholds we have reduced the number of cases in the contradictions from 12 to 2. This in fact raises the explanatory power by the reduced expressions, as only one case is not included in the solutions found for the cases with the outcome 0 as well as for the cases with the outcome 1. If we reduce the new data set, we find that the cases with the outcome 0 including the remainder cases are reduced to  $IKD\{0\}+IOD\{1\}FF\{2\}$ , meaning that all the cases with a low knowledge distribution (below 45%) broke down in the interwar Europe. This is the majority of breakdown cases (PL,GR,IT,PO,SP,HU,RO), The remaining two cases (AU,GE) are described by a high level of occupational distribution (>45%) and a high percentage of family farms.

On the other hand, the group of cases O(1) (Cases with the outcome 1) is reduced to  $IOD\{1\}FF\{0,1\}+IOD\{0\}IKD\{1\}FF\{2\}$ . The most cases (FR, NL, BE, UK) had a high occupational distribution combined with a low or medium percentage of family farms. The two cases FI and SW could not be reduced further, but their combination of a low occupational distribution, a high knowledge distribution and high percentage of family farms identifies them clearly as cases with the outcome 1.

### **3. Finding "good" thresholds**

In this paper the idea of the extension of QCA by Multi-Value Scales has been drafted. It is important to realise, that MVQCA unlike QCA uses non Boolean Data Representation, but that the two step approach of MVQCA is the same as the one of QCA: In the first step, the necessary prime implicants are calculated, and these are then combined to sufficient and necessary solutions in the second step. As the choice of thresholds is substantial for the success of the analysis some aspects on selecting thresholds will now be touched, which have to be discussed in the further development of the QCA method:

- One critique on QCA has been that the choice of thresholds can decide the result and therefore scientist can be tempted to adjust their thresholds so they gain the best results possible. This raises the question on the legitimacy of threshold manipulation as it is used above. Thresholds are used to separate data points representing phenomenons in the “real world”. These thresholds should be chosen so that they create natural subsets of all data points. This means that all data points have to be fitted into as homogenous subsets as

possible and that artificial cuts (e.g. between two near-by points) should be avoided. TOSMANA provides a “Thresholdsetter” – a visual aid for this task, where the thresholds can be set either by a calculated cluster subsection or be manipulated by hand. It is important that all thresholds should be set only looking at the data of the processed variable, and that the values of the outcome variable should have no influence at all on the threshold of the independent variables.

- It may appear strange to use a threshold of 45% instead of 50%. But it must be kept in mind that these values are all only numbers expressing a real world phenomenon. Due to the subject dealt with in social science, these are affected by a certain degree of uncertainty. This implies that 50% must not necessarily be a more appropriate threshold than 45%. Threshold are used to separate data from other data. Therefore thresholds should be chosen so that they are discriminating as much as possible (without being artificial), creating as homogenous data sets as possible. Therefore thresholds will often not to be convenient numbers (like 50%, 10.000€, ½ etc.).
- The used thresholds also have to be mentioned when using (MV)QCA results: Stating that cases are having a “low value with the Variable FF” does not mean the same with different thresholds. A threshold of 45% may produce a different set of subsets as a threshold of 50% (A case with 48% is below the threshold of 50% but above the threshold of 45%). Therefore, cases with a “low” value should correctly be addressed as “cases with a value below the threshold of 45% (or 50%) “ indicating the used threshold.
- It is possible to use as many thresholds as needed to remove all contradictions. But using too many thresholds will singulize cases so that all cases are represented by one combination at their own. Such data sets will not provide useful solutions, as it will be difficult to find parsimonious solutions within a set of singulized cases. The ideal threshold definition will use as grossly scaled variables as possible and explain as many cases as possible without contradictions. A dichotomous data set without contradiction with only a few number of variables would be the optimal data set for the further application of QCA results. The use of strictly dichotomious data sets will only be possible in few researches. But as the Vanhanen example given above shows, that it is possible to exclude all

contradictions possible to exclude<sup>17</sup>, by changing two thresholds and adding one more threshold to a variable. Therefore major advances seems possible with just a few wisely chosen steps.

#### **4. Conclusion**

Extending QCA with multi valued scales offers the opportunity to reduce the number of contradictions in data sets used for Boolean synthesis. But the example given above also shows the importance of the use of well chosen thresholds. In fact, only a good combination of a good number of coding possibilities and wisely selected thresholds can bring up a recoded data set, which can be applied successfully for Small-N comparative research.

One critical point of QCA applications is the choice of thresholds. In many applications of QCA the thresholds are chosen from a theoretical foundation, but the consequences of the choice are not checked, which may make the choice non defensible. A threshold set to a round number may appear reasonable, but if it cuts just through a number of closely related cases an artificial division is made, which is the source for questionable results. TOSMANA provides a couple of tools based on the ideas presented in this paper to find more suitable thresholds. But they can only be a first step in further discussions on how to set thresholds for the use with (MV)QCA. As mentioned in my introduction, faulty thresholds have been a major point of critique towards the QCA approach. Using more functional thresholds improves the use of (MV)QCA in the research process. Therefore the subject on choosing "good" thresholds should get more attention in further development of macro-qualitative comparative analysis tools.

---

<sup>17</sup> The two last cases could not be separated as their original data was too close to separate in a non-artificial way.

## 5. Bibliography

- Berg-Schlosser, Dirk and Gisèle DeMeur (1994): Conditions of Democracy in Interwar Europe: A Boolean Test of Major Hypotheses. In *Comparative Politics*, 26:4, 253-280.
- Berg-Schlosser, Dirk (2002): Macro-quantitative vs. macro-qualitative methods in the social sciences – testing empirical theories of democracy. Paper presented at the International Sociological Association World Congress, Brisbane.
- Brayton, Robert K. And Sunil P. Khatri (1999): Multi-value Logic Synthesis. Paper. University of Berkeley, CA. Paper Online in Internet: [http://www-cad.eecs.berkeley.edu/Respep/Research/mvsis/doc/mvsis\\_main.pdf](http://www-cad.eecs.berkeley.edu/Respep/Research/mvsis/doc/mvsis_main.pdf) [cited 2003-08-22].
- Cronqvist, Lasse (2002): How MVQCA works. A short Introduction to the Ideas of the Algorithm used in TOSMANA. Powerpoint-Presentation. Online in Internet: <http://staff-www.uni-marburg.de/~cronqvis/tosmana/resources/> [cited 2003-08-22].
- Cronqvist, Lasse (2003a): Tosmana – Tool for small-n analysis. Version 1.0. Marburg: University of Marburg. Online in Internet: <http://www.tosmana.net/> [cited 2003-08-22].
- Cronqvist, Lasse (2003b): Tosmana User Manual. Marburg: University of Marburg. Online in Internet: <http://www.tosmana.net/> [cited 2003-08-22].
- De Meur, Gisèle and Benoît Rihoux (2002): L'analyse quali-quantitative comparée (AQQC-QCA): approche, techniques et applications en sciences humaines. Louvain-La-Neuve.
- Drass, Kriss A. and Charles C. Ragin (1992): Qualitative Comparative Analysis 3.0. Evanston, Illinois: Institute for Policy Research, Northwestern University. Online in Internet: <http://www.fsqca.com> [cited 2003-08-22].
- Goldthorpe, John (1991): "Current Issues in Comparative Macrosociology." In *Comparative Social Research*, 16, 1-26.
- MVSIS Group: Multi-Valued Logic Synthesis. Software Release. Online in Internet: <http://www-cad.eecs.berkeley.edu/Respep/Research/mvsis/> [cited: 2003-08-22].
- Pennings, Paul (2003): Beyond dichotomous explanations: Explaining constitutional control of the executive with fuzzy-sets. In *European Journal of Political Research*, 42:4, 541-567.

Ragin, Charles C. (1987): *The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies*. Berkeley – Los Angeles – London: University of California Press.

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

Ragin, Charles C. and Kriss A. Drass (2002): *Fuzzy-Set/Qualitative Comparative Analysis 0.963*. Tucson, Arizona: Department of Sociology, University of Arizona. Online in Internet: <http://www.fsqca.com> [cited: 2003-08-21].

Vanhanen, Tatu (1984): *The Emergence of Democracy. A Comparative Study of 119 States, 1850-1979*. Commentationes Scientiarum Socialium 24. Helsinki: Finnish Society of Sciences and Letters.