R Manual for QCA (Version 4, July 2026)

Mello, Patrick A. (2021) *Qualitative Comparative Analysis: An Introduction to Research Design and Application*, Washington, DC: Georgetown University Press, *R* Manual for QCA (online appendix). https://doi.org/10.7910/DVN/KYF7VJ

The book is available at: http://press.georgetown.edu/book/qualitative-comparative-analysis

Introduction

The great (or perhaps infuriating) thing about R *is that there's always more to learn.*

Robert I. Kabacoff [1]

This *R* Manual for QCA complements *Qualitative Comparative Analysis: An Introduction to Research Design and Application* (Mello 2021). The focus is on preparing and analyzing data and conducting set-theoretic analyses, using the *R* software environment (R Core Team 2020), the RStudio text editor (2020), and we work primarily with the *R* packages *QCA* (Duşa 2019), *SetMethods* (Oana & Schneider 2018), and *ggplot2* (Wickham 2016). The *R* Manual is complemented by an *R* Script that contains the code used in this document (see below for download instructions).

As indicated in the above quote by Robert Kabacoff, one is hardly ever done learning *R*, though for new users the main hurdle is starting with *R* in the first place. This *R* Manual aims to *get you started* to help you conduct your own QCA study and reproduce the book's analytical steps. Unlike texts assuming prior knowledge of *R*, this text also covers basic questions of working with the *R* editor, shaping data, and issues that can arise along the way.

As for QCA functions, the manual covers one complete QCA *research cycle* as presented in the book (Mello 2021, 6), covering calibration, necessary conditions, truth tables, solution minimization, and result visualization. That said, this *R* Manual does not aim to replace a comprehensive introduction to *R* and its functionality, nor every way of working with the QCA packages. I recommend consulting a general introduction to *R* and data management. Several textbooks are available, some tailored to specific fields (e.g. Field et al.; Elff 2021; 2012; Gaubatz 2015; Imai 2017; Kabacoff 2015; Pollock & Edwards 2018). For further analytical functions, see the comprehensive treatment of the *QCA* package by Adrian Duşa (2019), and the guide to QCA using *R* by Nena Oana, Carsten Schneider, and Eva Thomann (Oana et al. 2021), with a complete discussion of the *SetMethods* package (Oana & Schneider 2018).

Another way of getting started with *R* is through the many free online resources, such as the *swirl* package.[2] Various online forums also answer all kinds of *R* questions, including on QCA-related packages.[3] When **errors occur**, a good first step used to be pasting the error message into a search engine, since others have often hit the same issue, especially for common tasks like installing *R* and its packages, reading data, or working with data frames. As AI models have become more powerful, these can often provide more efficient help with coding issues than search engines.

---

[1] Kabacoff (2015, 532).

[2] This can be reached at: https://swirlstats.com/.

[3] A general forum is: https://stackoverflow.com/.
There are also QCA discussion groups on Google (https://groups.google.com/g/qcawithr) and Facebook (https://www.facebook.com/groups/483487988377003/).

1 Introduction to *R*

Before we can start working, we first need to **install** *R* and RStudio. *R* is the software environment, and RStudio is an editor that enhances it with a graphical interface and integrated functions. We install *R* first and *only then* install RStudio (an open-source license usually suffices), since RStudio runs on top of *R*. The latest versions can be accessed at the following websites, which include platform-specific instructions for Windows, Mac OS X, or Linux:

R: https://cran.r-project.org/
RStudio: www.rstudio.com/products/rstudio/

One thing to keep in mind is that there are frequent updates for *R* and RStudio, and for many *R* packages, so you should regularly **check for updates**, especially after time away. Updates occasionally rename or discontinue functions or change default settings, though most updates do not disrupt existing code.

1.1 Online Material

Complementary online material can be found on my website https://patrickmello.com under the drop-down menu `QCA`. The files include an *R* Script (`RManualforQCA.R`) and data sets that are used throughout the following sections. This material is also available on Harvard Dataverse: https://doi.org/10.7910/DVN/KYF7VJ.

1.2 Getting to Know *R* and RStudio

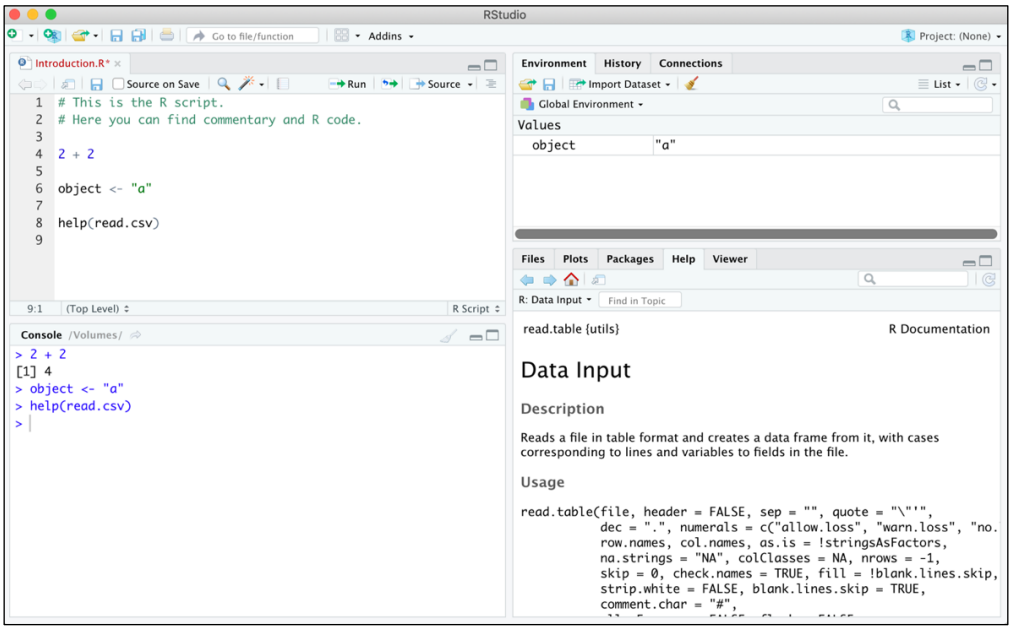
Once we have installed *R* and RStudio, we can start a new session by opening RStudio. It is not necessary to open *R*, because RStudio runs on the *R* installation. For all future steps we will always work from within RStudio.

When we open RStudio, we see four windows, similar to what is shown in *Illustration 1*. The top left window shows our `R SCRIPT` (to see a script, open the file `RManualforQCA.R`). The window on the lower left shows the `CONSOLE`. This area shows the output that *R* produces. Whenever we ask *R* to compute something, results appear in the `CONSOLE` window: printed in black, while blue font repeats the script's commands.

In the top right corner, we can see several tabs, including `ENVIRONMENT`. This area shows, among others, objects we created and data sets we read into *R*. Finally, in the lower right corner there is another window with several tabs including `HELP`. Here we can find documentation on *R* functions. There are many more functions, but the ones mentioned should do – to get us started.

You can reproduce the example in *Illustration 1* by creating a new script (select `NEW FILE` and then `R SCRIPT` from the drop-down menu `FILE`) and writing the commands shown. The first two lines are **commentary**, always preceded by a hash sign (#); anything after a hash on that line is ignored by *R*. RStudio highlights commentary in a different color to set it apart from code (by default, `green`, though like most things in RStudio, this can be customized under `PREFERENCES`).

Illustration 1: Layout of RStudio



Commentary is part of good coding practice: short comments help us remember what a piece of *R* code does and why, which matters when you revisit your own code months later, and makes it transparent and reproducible to others. Many academic journals now require the submission of *R* code along with the data used, so writing clear, documented code from the outset saves work later. It is also good practice to use white space and break long lines up for readability; RStudio can do this for you: select the code and choose the function `REFORMAT CODE` from the `CODE` rider in the menu.

The next few lines in *Illustration 1* include small examples of *R* code. For instance, you can use *R* to do mathematical calculations, the result of which will be shown in the console. Here, we write `2 + 2` and then `RUN` this code to get the result.

Running code can be done in two ways: select code and click the `RUN` button, or select the code and press `CONTROL + ENTER` (Windows) or `COMMAND + ENTER` (macOS), usually the faster option. With the cursor simply placed in a line (no selection), the *entire line of script* runs; select more code to run more, or select and run an entire *R* script.

For the remainder of this manual and unless stated otherwise, text with `GRAY BACKGROUND` is meant to be written in the upper left script window of RStudio (you also find the entire script in the `RManualforQCA.R` file). From RStudio, you can run the code and manipulate it to practice the different functions and arguments along the way.

The *R* language is based on **objects**. We create objects by using a leftward arrow sign. Once an object is created, it appears in the `ENVIRONMENT` shown in the top right window. For example, we can create the object `A`, which shall equal `2 + 2`:

```
A <- 2 + 2
```

We can also assign names to objects, by placing the characters in quotation marks:

```
City <- "Amsterdam"
```

Note that *R* is **case-sensitive**, so `City` and `city` are treated as different objects (calling the second object will return an error because we have not created it). In any event, you can always list everything in your environment with the *ls* function. This is followed by empty brackets, to produce the default function of *ls*, without any parameters or arguments:

```
ls()
```

This yields the following output in the console:

```
[1] "A" "City"
```

Note that the number in brackets `[1]`, shown in the output given in the console, is an **index number**. In this example, there are only two objects. But once we have more objects, the results may run over several lines, each of which would be preceded by an index number, so that separate objects can be clearly identified by their index number.

1.3 Working Directory

When starting to work with *R*, we specify a **working directory** from which *R* reads data and scripts, and where it will save any output we generate. To do this, we first check what the current working directory is, using the function *getwd*:

```
getwd()
```

This should return a pathname on your computer, ideally with all relevant files in it. We can check the folder content in our working directory with the function *list.files*:

```
list.files()
```

To change the working directory, we use the function *setwd*:

```
setwd("/mydirectory/mydata")  # Adapt this to match your own pathname
```

Note that *R* **does not recognize subfolders**, so all relevant files you want to work with in a single *R* session have to be in the same folder. Hence, it is a good idea to create a dedicated *R* project folder and place all data files and scripts there before working with *R*. Errors often come from the folder structure or from misspelled file names. In both cases, *R* will not find the right data file, and hence cannot read the data to begin with.

Should you run into trouble when setting your working directory, note that you can also do this manually in click-and-point fashion in RStudio. Simply navigate to the lower right window and the rider FILES (depending on your operating system, this may also be in another window). In the files view, you can steer to the desired folder, then select MORE and SET AS WORKING DIRECTORY. Use the *getwd* function from above to confirm your settings.

1.4 Installing Packages

R comes with a multitude of different functions. Yet, one of the great advantages of *R* is the existence of user-defined **packages** that include functions and code for specific purposes. Rather than writing a long patch of code each time we seek to conduct a sequence of analytical steps, a function from a specific package may do just that with a single command.

To work with packages, we first need to install them on our local computer system. For our purposes, we mainly work with the packages *QCA* (Duşa 2019), *SetMethods* (Oana & Schneider 2018), and *ggplot2* (Wickham 2016). We install the three together, using *install.packages* and the following syntax (we could also run separate commands for each package):

```
install.packages(c("QCA", "SetMethods", "ggplot2"))
```

Occasionally, *install.packages* yields an **error message** stating that the package is "not writable" on the system library (because your account may not have write permissions on the computer) and asking whether you would like to use a "personal library" instead. In that case, you can type "yes" to proceed and install the package. You can also solve the issue by changing the permissions settings on your computer or by using a different file directory.

While packages only need to be installed once, they must be **loaded at the start of each session**, using the *library* function. To load all three packages, we use the following:

```
library("QCA")
library("SetMethods")
library("ggplot2")
```

Note that you can always check which packages are installed and loaded in point-and-click fashion, by navigating the cursor to the rider PACKAGES in the lower right window. There you find INSTALL and UPDATE buttons, which may be more convenient than running code to achieve the same aims.

For summary information on your *R* environment, run *sessionInfo*. This gives you info on your *R* version, platform and operating system, general settings, and attached packages and their version numbers:

```
sessionInfo()
```

1.5 Citing Packages

Note that *R* and *R* packages, like other publications, should be **cited** when used in your own work. To retrieve the full reference, use *citation* with empty brackets (on how to cite the *R* version), and together with the package name to cite specific packages:

```
citation()
citation("QCA")
citation("SetMethods")
citation("ggplot2")
```

For the first line, this yields the following information:

```
To cite R in publications use:
R Core Team (2026). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
doi:10.32614/R.manuals <https://doi.org/10.32614/R.manuals>.
<https://www.R-project.org/>
```

2 Data Management

2.1 Reading Data

To work with *R*, we need to import our data first. *R* can read various file formats, but for QCA, the most common format is comma-separated files (.csv). Here, we use the function *read.csv* to read the file QCAdata.csv (part of the online material referred to above) and create a new data frame that we choose to call QCA.data (we could assign any other name). The data then appears in our Environment in the upper right corner of RStudio.

```
QCA.data <- read.csv("QCAdata.csv")
```

However, when we examine the new QCA.data object by running its name (or by clicking on it in the environment), we can see that the code from the previous line has not read the data correctly. All the information has been read as a single column because the column separator has not been recognized. We can solve this problem by telling *R* that our data is separated by semicolons rather than commas, using the sep argument:

```
QCA.data <- read.csv("QCAdata.csv", sep = ";")
```

The result looks better, but the case-name column has been treated as a variable. We fix this with the *row.names* argument, specifying which column holds case names (usually the first). With this change we get 35 cases and 4 conditions in our environment (A, B, C, and OUT), and the row-name column is shaded gray to show it is not treated as a variable:

```
QCA.data <- read.csv("QCAdata.csv", sep = ";", row.names = 1)
```

The *read.csv* function can be further customized: *row.names* specifies the column with case names, header whether the first row holds variable names, sep the separator used (comma, semicolon, space), and dec the decimal sign (comma or period). These are not always needed, since the defaults often read the data correctly, but they are the first things to check if you run into errors. For more, consult the documentation:

```
?read.csv
```

2.2 Pre-Existing Data Sets

Some packages already contain data sets that may be useful for exercises. For example, let us take a look at the data entailed in *SetMethods* (Oana & Schneider 2018):

```
data(package="SetMethods")
```

This yields a list of data sets, which we can load with the *data* function. To explore, we load the data from a QCA study by Barbara Vis (2009). The data will appear in our environment:

```
data(VISF)
```

To examine this data set, we can either click on the object in the environment (this will open a new rider) or simply type its name, which will yield the entire data set, printed in the console. But this may not be convenient for large data sets. To simply inspect the structure and see whether the data was read correctly into *R*, we use the *head* function, which yields the first six lines from the data set, as shown below the command `head(VISF)`. We see that the data set contains government cabinets as cases, and four fuzzy-set conditions (`u` is the outcome):

```
head(VISF)
               p    s   r    u
Lubbers1    0.33 0.83 1.0 0.83
Lubbers2    0.17 0.33 1.0 0.33
Lubbers3    0.33 0.67 0.6 0.67
Kok1        0.17 0.40 0.4 0.67
Kok2        0.33 0.33 0.4 0.17
Balkenende2 0.67 0.67 1.0 0.83
```

2.3 Creating New Data Sets

We can also **create new data sets** with *R*. For a simple data set, we start by creating several vectors. These are one-dimensional arrays with numeric, character, or logical data. For a start, we only need numerical data, which shall represent raw data values for our cases. We use the *combine* function `c()` to create vectors with three values, separated by commas. We create the vectors `a`, `b`, and `v3` and assign values to each (with decimals for vectors `b` and `v3`):

```
a <- c(1, 3, 17)
b <- c(24.1, 18.0, 10.3)
v3 <- c(6.34, 3.78, 0.61)
```

Vectors represent **columns** in our QCA data frame (cases' values for a specific condition). The vectors can be tied together with the *data.frame* function:

```
our.data <- data.frame(a, b, v3)
```

To work with the data, we should first assign **row names**, because the rows represent our cases:

```
row.names(our.data) <- c("Shanghai", "Karachi", "Seoul")
```

We can also assign **new column names** (instead of the generic `a`, `b`, and `v3`) by creating a new character vector and placing it inside our data frame. These are the conditions that we use in QCA (including the outcome condition):

```
columns <- c("rank", "population", "size")
colnames(our.data) <- columns
```

Let's examine our new data set by typing `our.data` in the console:

```
our.data
         rank population    size
Shanghai    1       24.1    6.34
Karachi     3       18.0    3.78
Seoul      17       10.3    0.61
```

We can see that we have created three cases (Shanghai, Karachi, and Seoul) and three conditions (rank, population, and size), the starting point for a larger data set for QCA. We would still need to calibrate it, since it is not yet in crisp or fuzzy-set format. Most QCA studies import their data from Excel and the like rather than creating it from scratch in *R*, but the code above shows how, if needed. Next, let's see how data can be modified.

2.4 Modifying Data Sets

Elements in a data frame can be modified. This may be needed, for instance, when we decide to **drop a case** or condition, or when we want to **assign a new value** to a case. **Modifying column names** can be done with the function *colnames* and a number that specifies the column we want to change. This number is placed in square brackets. For example, we can apply *colnames* to `our.data` and change the name of column two by assigning a new name to it. Note that this will overwrite any existing name for the respective column:

```
colnames(our.data)[2] <- "new name"
```

On other occasions, you may want to **add a case**, which can be done by attaching a vector to the existing data frame as an additional row, using the *nrow* argument and specifying its values after the *list* argument (here we choose the values 27, 8.7, and 1.57):

```
our.data[nrow(our.data) + 1,] = list(27, 8.7, 1.57)
```

We **assign a name** to the new case using the *row.names* function and specifying the data set (`our.data`), the row number (in square brackets), and by inserting the name of the new case:

```
row.names(our.data)[4] <- "London"
```

Sometimes we may want to assign **new values to specific cells** in the data frame. For example, we can attribute a value of 11.2 to the cell in row 4 and column 2, which is London's value in the condition represented by the second column (the one we relabeled to `new name`). To do this, we specify the data frame (`our.data`) and use square brackets to indicate the value we want to change. Inside the square brackets, the first number refers to the rows and the second number refers to columns. These are always separated by a comma. We specify the target cell in this way and assign a new value of 11.2 to it:

```
our.data[4,2] <- 11.2
```

Finally, we can **delete rows and columns**. To delete a row, we create a subset of our data (placing the result in a new data frame) using a minus sign inside square brackets with the row number. Leaving the space after the comma empty selects all columns, so this creates a new data frame without row 2 and its values across every column:

```
new.data <- our.data[-2,]
```

Deleting a column from a data frame can be done in the same manner, for instance to create a subset that excludes columns 2 to 3 (or any other range of columns). For this to work, we need to place the column range in regular brackets, inside the square brackets:

```
new.data <- our.data[,-(2:3)]
```

Another way to delete a column is by stating its number in square brackets and assigning a value of `NULL` to it. For example, to delete the second column and all respective rows from `our.data` we can use the following code:

```
new.data[,2] <- NULL
```

These examples help address some basic questions that arise when working with QCA data frames. That said, with *R* there are always multiple ways how to solve an issue and there may be more economical ways to achieve the same aims in your *R* script, especially when manifold operations are required on larger numbers of cases. Such options are explored in greater detail, for instance, in Field, Miles & Field (2012), Kabacoff (2015), and Elff (2021).

2.5 Saving Data Sets

The old adage *save early, save often* also holds true for *R*. Closing RStudio prompts us to save changes to our *R* script, and we can save our workspace to pick up where we left off. We should also save our data regularly, which just requires naming the object to save. Here, we decide to **save the data frame** we created earlier (`our.data`) as a new comma-separated file `mydata.csv`, using the *write.csv* function. Running this line creates the new file in our working directory (check with the *list.files* function):

```
write.csv(our.data, "mydata.csv")
```

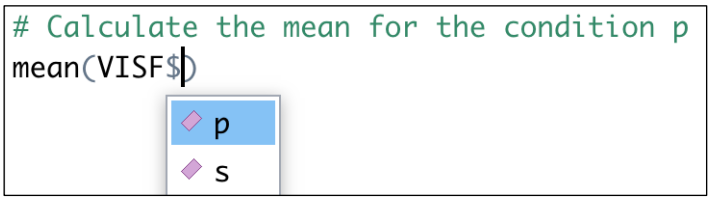
3 Descriptive Statistics

3.1 Descriptive Measures

Before engaging in set-theoretic analysis, it is useful to examine data with descriptive statistics. We can use the Vis (2009) data entailed in the *SetMethods* package (see above). To **select a column** within a data frame, we use the `$` operator. For example, to get the *mean* for the condition `p` and the *median* for condition `s` of the `VISF` data set, we use the code below. RStudio shows the available conditions that can follow the dollar sign as you type (see *Illustration 2*). This helps prevent misspellings and shows the contents of complex objects (used again below):

```
mean(VISF$p)
[1] 0.354
median(VISF$s)
[1] 0.4
```

Illustration 2: Dollar Sign in RStudio



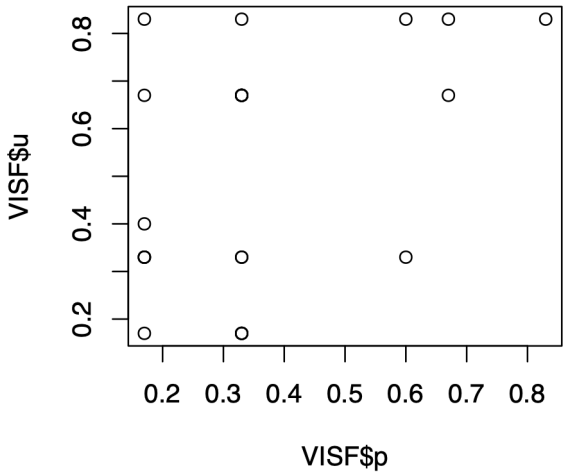
It is also useful to obtain summary statistics for an entire data set at once, using the *summary* function. This returns the minimum, maximum, quartiles, and mean for every column:

```
summary(VISF)
```

We can also **visualize** the relationship between conditions by plotting one against the other (XY plot), using the basic *plot* function of *R*. All we need to do is specify two conditions in the data frame, using the syntax introduced earlier. The output will appear in the PLOTS window on the lower right of RStudio (*Illustration 3*):

```
plot(VISF$p, VISF$u)
```

Illustration 3: XY Plot



To save the plot as a file on our computer, we can select EXPORT at the top of the PLOTS window and specify the file type, format, and resolution. Another way to save plots is to include such a function directly in our script. For instance, to save our plot as a PDF file, we run the following code. The first line calls upon *R*'s graphics device driver to produce PDF files, specifying the name of the file we want to create. The second line is identical to the code we used before. Finally, the command *dev.off* closes the pdf graphics device, so that the file is created:

```
pdf(file = "XYplot.pdf")
plot(VISF$p, VISF$u)
dev.off()
```

Like most functions, the PDF graphics device can be customized in many ways (e.g., to specify *width*, *height*, and *fonts* used). How this is done can be seen in the documentation:

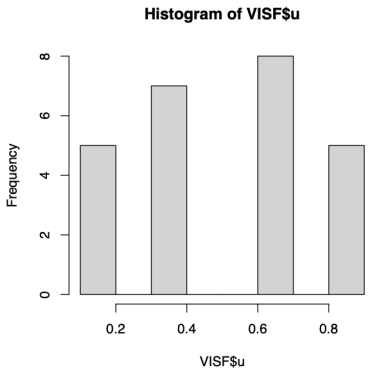
```
help("pdf")
```

A basic XY plot serves to check the relationship between a condition and the outcome. It can also be used to plot raw data against calibrated data. While the *plot* function can be tweaked to a certain extent, for more refined visualizations and different kinds of plots, I recommend the *ggplot2* package (Wickham 2016).[4] Examples are provided in later sections of this *R* Manual.

Histograms are another helpful descriptive tool. For any condition, a histogram shows the distribution of values in the data. Histograms are created with the *hist* function, where we need to specify the condition we seek to examine. The result is shown in *Illustration 4*:

```
hist(VISF$u)
```

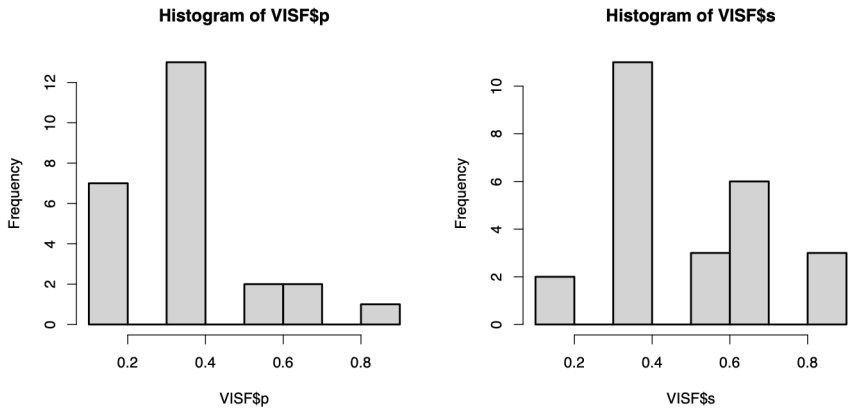
Illustration 4: Histogram



The *hist* function can be customized. Suppose we want to place several histograms on the same page, to document them in the appendix to our QCA study. This can be achieved with the *par* function to customize graphical parameters in *R*. The argument *mfrow* specifies the number of rows and columns. In this case, we want to place two histograms on one row with two columns (*Illustration 5*). The argument *lwd* specifies line width (thicker containers):

```
par(mfrow = c(1, 2), lwd = 2)
hist(VISF$p)
hist(VISF$s)
```

Illustration 5: Histograms, Side by Side



[4] The *R Graph Gallery* has a large collection of charts done with *R*: https://www.r-graph-gallery.com/.

3.2 Histogram (Template)

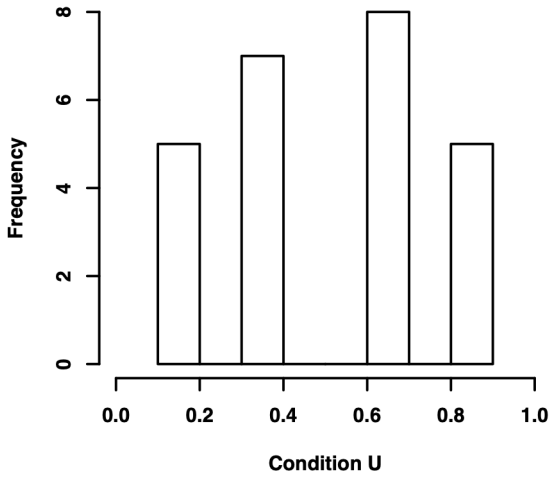
Finally, we may want to further **customize histograms**. For example, the following code produces white containers (`col = "white"`), an x-axis from 0 to 1 (`xlim = c(0, 1)`), an adapted label for the x-axis (`xlab = "Condition U"`), same font size for the axis labels and the axis numbers (`font = 2, font.lab = 2`), a solid line type (`lty = 1`), increased line thickness for the axes [`lwd = 2`], and an empty title [`main = ""`]. The result is shown in *Illustration 6*. In the same manner, we could also adapt the code to produce several histograms and plot them on the same page:

```
hist(
  VISF$u,
  col = "white",
  xlim = c(0, 1),
  xlab = "Condition U",
  font = 2,
  font.lab = 2,
  lty = 1,
  lwd = 2,
  main = ""
)
```

To reset the graphical parameters to a single column and row, we can use the following:

```
par(mfrow=c(1, 1))
```

Illustration 6: Histogram, Customized



The layout in *Illustration 6* works well for QCA purposes. Hence, we could use it as our standard layout to document the data distribution for all conditions and the outcome (ideally all histograms should be shown on a single PDF page). Note that in the above script, the x-axis is limited to scores between 0 and 1. When raw data is displayed, then the limits would need to be adjusted accordingly, similarly for the y-axis (depending on the frequencies in the data). For an example from a published QCA study, see the supplemental file to Meissner & Mello (2022): https://doi.org/10.1080/13523260.2022.2059226

4 Set-Theoretic Operations

The following sections describe the *R* functions needed to reproduce the analytical steps discussed in Mello (2021). The functions require that you have installed and loaded the packages *QCA* (Duşa 2019) and *SetMethods* (Oana & Schneider 2018).

4.1 Boolean Operators

For **Boolean operations** with crisp and fuzzy sets, we use the *pmin* and *pmax* functions of base *R*. These reflect the calculation of the minimum (Boolean AND) and the maximum (Boolean OR). To negate set membership scores, we simply use subtraction (Mello 2021, 49-53).

We first create a data frame `Boole` with the information from Table 3.3 (Mello 2021, 51):

```
A <- c(1, 1, 0)
B <- c(0, 1, 0)
C <- c(0.9, 0.7, 0.2)
D <- c(0.3, 0.8, 0.4)
Boole <- data.frame(A, B, C, D)
```

For the **Boolean AND** we use the *pmin* function across the respective conditions. We place the results into new conditions of the data frame:

```
Boole$AandB <- pmin(A, B)
Boole$CandD <- pmin(C, D)
```

Likewise, for the **Boolean OR** we use the *pmax* function across the respective conditions, also placing the results into new conditions:

```
Boole$AorB <- pmax(A, B)
Boole$CorD <- pmax(C, D)
```

Finally, for **Boolean negation** (NOT), we use subtraction, creating further conditions:

```
Boole$notA <- 1-A
Boole$notC <- 1-C
```

This results in a data frame that reflects all the calculation of Table 3.3 (Mello 2021, 51):

```
Boole
  A B   C   D AandB CandD AorB CorD notA notC
1 1 0 0.9 0.3     0   0.3    1  0.9    0  0.1
2 1 1 0.7 0.8     1   0.7    1  0.8    0  0.3
3 0 0 0.2 0.4     0   0.2    0  0.4    1  0.8
```

4.2 Compute Function

The *QCA* package (Duşa 2019) includes a *compute* function that evaluates Boolean expressions directly, without requiring us to manually combine *pmin*, *pmax*, and subtraction as shown above. This becomes especially useful once we work with more complex expressions, such as

those returned by the *minimize* function in later chapters. For instance, using the VISF data set introduced earlier, we can create a new condition that reflects membership in the Boolean expression p*s*~r (the intersection of p and s, excluding r):

```
data(VISF)
VISF$newcondition <- compute("p*s*~r", data = VISF)
```

The expression syntax follows standard QCA notation: an asterisk (*) for the logical AND, a plus sign (+) for the logical OR, and a tilde (~) to negate a condition. We will use the *compute* function extensively in the chapter on visualizing results, where it lets us calculate membership in a solution term directly from the Boolean expression returned by minimize.

4.3 Venn Diagrams

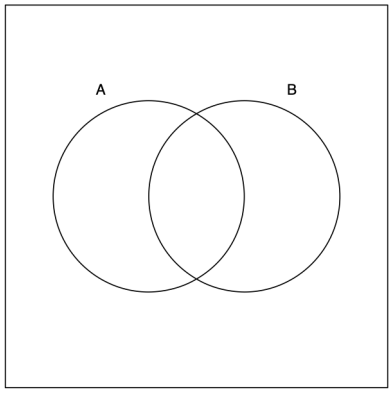
Set operations on crisp sets can be visualized with **Venn diagrams**. The *venn* package (Duşa 2022) allows the drawing of Venn diagrams with up to 7 sets. This package is installed as part of the *QCA* package (Duşa 2019), so usually we wouldn't need to install it separately. In any event, we install and load the package:

```
install.packages("venn")
library("venn")
```

For a simple Venn diagram, we only need to specify the number of sets to be included. The result is show in *Illustration 7*:

```
venn(2)
```

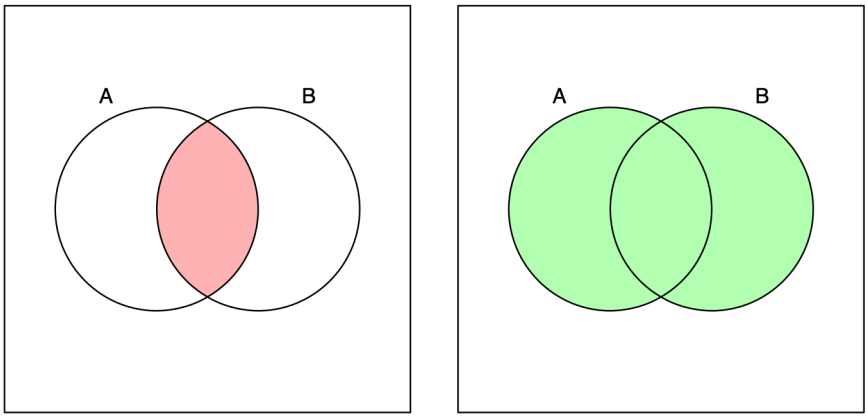
Illustration 7: Venn Diagram with Two Conditions



We can also create Venn diagrams where the intersection and the union are highlighted, as show in *Illustration 8* (we place them side by side with the *par* function introduced above):

```
par(mfrow = c(1, 2))
venn("A*B", zcol = "red")
venn("A+B", zcol = "green")
```

Illustration 8: Venn Diagrams (Intersection and Union)



5 Calibrating Sets

5.1 Creating Raw Data

Before we can calibrate sets, we need raw data to work with. To illustrate the procedure, we create a data frame with random data for our calibration. We use the function *runif* to draw from a uniform distribution 30 random numbers between 0 and 60 and place these inside the vector `Raw1`. We repeat this for `Raw2` and then bind the vectors together in the data frame `DT`:

```
Raw1 <- runif(30, min = 0, max = 60)
Raw2 <- runif(30, min = 0, max = 60)
DT <- data.frame(Raw1, Raw2)
```

For convenience, and also since we do not need precision beyond two **decimal scores**, we round up the data frame to two digits:

```
DT <- round(DT, digits = 2)
```

Now we can examine the data frame with the *head* function, which returns the first six lines (note that you will get **different values** on your computer because these were randomly created):

```
head(DT)
Raw1  Raw2
1 11.00 13.56
2  2.48 51.07
3  3.50  3.88
4 48.51 56.38
5 31.80 33.35
6  4.35 36.92
```

5.2 Direct Method of Calibration

To calibrate a new crisp or fuzzy set, we need a data frame with **raw data**. This can be transformed using the *calibrate* function of the *QCA* package (Duşa 2019), as shown in the next example. In brackets, we specify the raw data column of the data frame, the type of target

set (`crisp` or `fuzzy`), the method of calibration (`direct` or `indirect`), and the empirical anchors for full non-membership, the cross-over, and full membership in the target set. These are termed `thresholds` in the *QCA* package. Note that the thresholds are always stated in the same sequence, starting with the value that reflects being *fully outside* the target set.

For example, we can use the `DT` data frame we created in the previous section to calibrate the raw data for condition 1 (`Raw1`), into a fuzzy-set condition `Fuzzy1`, using the thresholds 10, 30, and 50 (the latter reflects *full set membership*). This is the **direct method of calibration**:

```
DT$Fuzzy1 <- calibrate(DT$Raw1, type = "fuzzy",
                       method = "direct", thresholds = "10, 30, 50")
```

For presentational purposes, it may be helpful to use the **rounding function** again after the fuzzy calibration (see previous), to round to two decimals. This suffices to have a meaningful differentiation and more fine-grained scores yield no additional benefit for the analysis.

We can also calibrate a **crisp set** based on the same data and approach. For this, we only need to specify the cross-over point that distinguishes exclusion from inclusion:

```
DT$Crisp1 <- calibrate(DT$Raw1, type = "crisp",
                       method = "direct", threshold = "30")
```

5.3 Fuzzy-Value Assignment

Alternatively, we can also use a **qualitative approach** (of sorts), where we assign fuzzy values to specific ranges of scores in the raw data. This procedure is more complicated because we need to proceed stepwise, starting by creating an empty vector. This serves as a "container" to be filled with values throughout the calibration process:

```
Qual1 <- NA
```

Now we start by assigning values to cases that pass the threshold for *full inclusion*. We refer to the raw data column we used before and assign a value of 1 to all cases that have a raw data of 50 or higher.

```
Qual1[DT$Raw1>=50] <- 1                  # Full inclusion
```

In the next steps, we fill the vector sequentially with values for being *more in than out*, the *cross-over*, *more out than in*, and the threshold for being *fully outside* the set (full exclusion). The arrow on the right side indicates which fuzzy values are to be assigned to which range of raw data values:

```
Qual1[DT$Raw1<50 & DT$Raw1>30] <- 0.7    # "More in than out"
Qual1[DT$Raw1==30] <- 0.50               # Cross-over
Qual1[DT$Raw1<30 & DT$Raw1>10] <- 0.3    # "More out than in"
Qual1[DT$Raw1<=10] <- 0                  # Full exclusion
```

Finally, we add the new vector to the data frame – creating a column with the same name – and have a look at the data. Again, your numbers will differ because these were drawn randomly:

```
DT$Qual1 <- Qual1
head(DT)
  Raw1  Raw2 Fuzzy1 Crisp1 Qual1
1 32.73 24.90   0.60      1  0.7
2 13.69 19.77   0.08      0  0.3
3 15.54 43.81   0.11      0  0.3
4 26.92 45.25   0.39      0  0.3
5  1.80 26.08   0.02      0  0.0
6 43.23 30.55   0.88      1  0.7
```

We can now compare the calibration approaches used to create three sets from the `Raw1` data. Qualitatively, `Fuzzy1`, `Crisp1`, and `Qual1` all fall on the same side of the 0.50 cross-over. Numerically they differ, though, because each uses a different scale (two values for the crisp set, five for the qualitative fuzzy set, and continuous values for the direct method), as discussed in Chapter 5 (Mello 2021).

It is also worth clarifying how these two approaches map raw scores onto fuzzy values. The qualitative approach above assigns a small set of fixed values (0, 0.3, 0.5, 0.7, 1) to entire ranges of raw scores, producing a step-wise calibration: cases within the same range receive an identical fuzzy score. The direct method, by contrast, fits raw scores onto a continuous logistic (S-shaped) curve anchored at the three thresholds, so cases receive distinct fuzzy scores that vary smoothly with their distance from the thresholds. It therefore preserves more fine-grained information, whereas the qualitative approach may suit cases where only ordinal distinctions between ranges are analytically meaningful.

5.4 Threshold Finder

Identifying appropriate thresholds is one of the more delicate steps in calibration. The *findTh* function of the *QCA* package automatically proposes thresholds for a given number of anchors, based on gaps in the distribution of the raw data. This can be a useful starting point, but it should never replace substantive reasoning about where the meaningful breakpoints in the data lie. This is what we receive when asking for three thresholds (full exclusion, cross-over, and full inclusion) for the Raw1 data we created above:

```
findTh(DT$Raw1, n = 3)
[1] 21.595 31.810 48.360
```

Because Raw1 is generated anew each time with *runif*, the exact threshold values above are specific to this run and will not reproduce on your own computer. What matters is the pattern: *findTh* looks for natural gaps in the distribution of the raw data and proposes anchors close to those gaps, so your own output will differ in the precise numbers but follow the same logic.

5.5 Checking for Ambiguous Cases

After calibration, it is good practice to check whether any cases were assigned a fuzzy value of exactly 0.5, since these cases are neither more in nor more out of the target set and can

complicate the analysis. For this, we can use the *ambig.cases* function of the *SetMethods* package, to flag any such cases for the calibrated condition Fuzzy1:

```
ambig.cases(DT$Fuzzy1)
[1] "There are no cases with fuzzy-set scores of 0.5."
```

For our random data, none of the cases land exactly at the cross-over point, so no further adjustment is needed before moving on to the analysis. Had any cases turned up here, it would be worth revisiting the calibration thresholds or examining those cases more closely before proceeding.

5.6 Skewness Check

Calibrated sets should also be checked for skewness, meaning a strong imbalance between the number of cases more in and more out of the set. Substantial skew can affect consistency and coverage scores in ways that are easy to overlook. The *skew.check* function of the *SetMethods* package reports this for one or more calibrated conditions at once:

```
skew.check(DT[, 3:5])
[1] "Set Fuzzy1 - Cases > 0.5 / Total number of cases: 16 / 30 = 53.33 %"
[2] "Set Crisp1 - Cases > 0.5 / Total number of cases: 16 / 30 = 53.33 %"
[3] "Set Qual1 - Cases > 0.5 / Total number of cases: 16 / 30 = 53.33 %"
```

All three calibrated conditions show an identical, close-to-even split here, with 16 of the 30 cases above the cross-over point in each. A result this balanced points to little skewness for this particular draw of raw data. A much more lopsided split (say, 27 out of 30 cases on one side) would be a signal to reconsider the calibration before treating the set as analytically meaningful.

6 Necessary Conditions

There are different ways to assess necessary conditions, but the recommended function for most purposes is *QCAfit* from the *SetMethods* package (Oana & Schneider 2018). It tests whether a specified number of conditions, in their absence or presence, are individually necessary for the outcome. We start by reading the QCAdata.csv file from the online material:

```
QCA.data <- read.csv("QCAdata.csv", sep = ";", row.names = 1)
```

For the *QCAfit* function we need to specify the columns with our calibrated conditions in square brackets (here columns 1 through 3), the outcome (QCA.data$OUT), that we want to test for necessity (we could also test for sufficiency by setting this to FALSE), and that we want to analyze the outcome rather than the non-outcome, hence the negative outcome is set to FALSE:

```
QCAfit(QCA.data[, 1:3], QCA.data$OUT, necessity = TRUE, neg.out = FALSE)
   Cons.Nec Cov.Nec   RoN
A     0.566   0.848 0.915
B     0.639   0.955 0.973
C     0.761   0.645 0.582
```

```
~A    0.606   0.552 0.576
~B    0.538   0.491 0.545
~C    0.389   0.666 0.858
```

None of the three conditions or their negations reaches the conventional 0.90 consistency threshold for necessity with respect to the outcome; the closest is condition C at 0.761, still short of that mark (on the customary thresholds for consistency, coverage, and relevance of necessity, see Mello 2021, Chapter 6). In the next step, we analyze the non-outcome, setting the respective parameter to TRUE:

```
QCAfit(QCA.data[, 1:3], QCA.data$OUT, necessity = TRUE, neg.out = TRUE)
   Cons.Nec Cov.Nec   RoN
A     0.358   0.409 0.736
B     0.270   0.309 0.703
C     0.745   0.482 0.489
~A    0.867   0.604 0.606
~B    0.960   0.670 0.649
~C    0.452   0.591 0.832
```

Here, the absence of B (~B) stands out at 0.960, comfortably above the threshold, with a coverage of 0.670 and a relevance of necessity of 0.649. This suggests that the absence of B comes close to being a necessary condition for the non-outcome, a pattern worth keeping in mind once we turn to the truth table and the derivation of solution terms below.

7 Truth Table Analysis

We create truth tables with the *truthTable* function of the *QCA* package (Duşa 2019), specifying the data frame (QCA.data), the outcome ("OUT"), whether to include logical remainder rows (TRUE), whether to show cases (TRUE), and the consistency cutoff (typically 0.75 or higher) that determines which rows enter the minimization. Finally, we sort the table by consistency ("incl") and case count ("n"):

```
TT <- truthTable(QCA.data, "OUT", complete = TRUE,
                 show.cases = TRUE, incl.cut = 0.75,
                 sort.by = "incl, n")
```

Once created, we can examine the truth table by calling upon the object:

```
TT
  OUT: output value
    n: number of cases in configuration
 incl: sufficiency inclusion score
  PRI: proportional reduction in inconsistency
    A  B  C    OUT    n   incl  PRI
4   0  1  1     1     9   0.995 0.991
6   1  0  1     1     4   0.994 0.989
7   1  1  0     1     6   0.944 0.916
5   1  0  0     0     4   0.594 0.165
```

```
2   0  0  1     0     12  0.497 0.239
1   0  0  0     ?     0     -     -
3   0  1  0     ?     0     -     -
8   1  1  1     ?     0     -     -
(Case identifiers for each configuration have been omitted here for space;
show.cases = TRUE in the code above still returns them.)
```

Three configurations pass the 0.75 consistency threshold and are coded as leading to the outcome: rows 4 and 6 are highly consistent (0.995 and 0.994) and each cover a handful of cases, while row 7 is somewhat less consistent (0.944) but covers six cases on its own. Two configurations fall below the threshold and are coded as not leading to the outcome, and three configurations have no empirical cases at all (marked "?"); these logical remainder rows will become relevant once we derive the parsimonious and intermediate solutions below.

This code assumes that we are working with a data frame that only includes calibrated conditions and an outcome. But sometimes we may want to specify the conditions to be included. For an alternative truth table (TT2), we can use the conditions argument and a vector that lists all the conditions we want to include:

```
TT2 <- truthTable(QCA.data, "OUT", complete = TRUE,
                  show.cases = TRUE, incl.cut = 0.75,
                  conditions = c("A", "B"),
                  sort.by = "incl, n")
TT2
  OUT: output value
    n: number of cases in configuration
 incl: sufficiency inclusion score
  PRI: proportional reduction in inconsistency
    A  B    OUT    n   incl  PRI
2   0  1     1     9   0.987 0.978
4   1  1     1     6   0.956 0.931
3   1  0     1     8   0.798 0.679
1   0  0     0     12  0.456 0.197
```

Restricting the truth table to A and B only collapses the rows that previously differed by C, which is why the case counts and consistency scores here differ from those in the full truth table above. Three of the four possible combinations of A and B pass the 0.75 threshold, and only the combination where both conditions are absent (row 1) falls short, at 0.456.

Once created, we can examine the truth table by calling the object (TT or TT2), or extract the truth table proper to save separately: the object tt (lowercase) sits inside the larger truth table object, which also holds additional information the *R* package needs. You can always explore an object's structure with *R*'s *str* function.

```
TT                          # Display the truth table TT
str(TT)                     # Display the structure of the TT object
write.csv(TT$tt, "TT.csv") # Save the truth table element in a new file
```

8 Solution Types

We derive solution terms from the truth table with the *minimize* function of the *QCA* package (Duşa 2019). This function applies the rules of Boolean minimization to the rows that consistently lead toward the outcome, that is, the rows that pass the consistency threshold we defined when we created the truth table (incl.cut above).

8.1 Conservative Solution

The minimization can be customized, but requires a truth table object (TT) and a choice of which rows to include: "1" for consistent rows only, or "?" to also include logical remainders. Setting details and use.tilde to TRUE displays measures of fit and uses the tilde sign (~), rather than lowercase, for absent conditions. This yields the conservative solution:

```
sol.cons <- minimize(TT, include = "1", details = TRUE, use.tilde = TRUE)
```

Let us look at this solution by calling upon the object sol.cons:

```
sol.cons   # Examine the solution
M1: A*B*~C + A*~B*C + ~A*B*C -> OUT
           inclS   PRI   covS   covU
-------------------------------------
1  A*B*~C  0.944  0.916  0.270  0.179
2  A*~B*C  0.994  0.989  0.312  0.167
3  ~A*B*C  0.995  0.991  0.397  0.239
-------------------------------------
       M1  0.978  0.970  0.746
```

The conservative solution consists of three sufficient paths, none reducing to a single condition: A*B*~C, A*~B*C, and ~A*B*C. Each is highly consistent (all above 0.94), and together they cover most cases with the outcome, at an overall coverage of 0.746. Because this solution draws only on empirical rows without relying on logical remainders, it tends to be the most cautious but also the most complex of the three solution types.

8.2 Parsimonious Solution

We can also derive the parsimonious solution. This requires a complete truth table, meaning one that includes logical remainder rows. The main difference lies in the include argument, which now indicates that logical remainders should be considered ("?"):

```
sol.pars <- minimize(TT, include = "?", details = TRUE, use.tilde = TRUE)
```

As expected, the parsimonious solution differs from the conservative solution:

```
sol.pars
M1: B + A*C -> OUT
        inclS   PRI   covS   covU
---------------------------------
1    B  0.955  0.937  0.639  0.443
```

```
2  A*C  0.994  0.991  0.362  0.167
----------------------------------
    M1  0.964  0.953  0.805
```

Where the conservative solution needed three conjunctural paths, the parsimonious solution reduces to two simpler paths: B on its own, and A*C. The single condition B already covers both groups of cases from the conservative solution's first and third paths, which is only possible because the algorithm is now free to draw on logical remainder rows as simplifying assumptions. Consistency (0.964) and PRI (0.953) are close to the conservative solution, while coverage rises to 0.805, illustrating the usual trade-off between complexity and coverage.

To see which logical remainder rows were used as simplifying assumptions (SA) for the parsimonious solution:

```
sol.pars$SA       # Logical remainder rows 3 and 8 were used as SAs
$M1
  A B C
3 0 1 0
8 1 1 1
```

Rows 3 (~A*B*~C) and 8 (A*B*C) were treated as simplifying assumptions: the algorithm assumed that these hypothetical configurations, which have no case in our data, would also lead to the outcome.

8.3 Intermediate Solutions

Finally, we can derive the intermediate solution from the truth table. As discussed in Mello (2021, Chapter 7), there are two ways to create this type of solution: we can tell the algorithm to exclude certain logical remainder rows from the minimization, those we deem implausible on substantive or logical grounds, or we can formulate directional expectations that the algorithm then uses when selecting among logical remainders.

To make an informed choice, let us look again at the complete truth table above: rows 1, 3, and 8 are logical remainders (marked "?" in TT). Row 3 (~A*B*~C) has no cases in that particular configuration; we may deem it implausible on substantive grounds and choose not to include it as a counterfactual.

To exclude this single row, we use the exclude argument:

```
sol.int1 <- minimize(TT, include = "?", details = TRUE, use.tilde = TRUE,
                     exclude = c(3))
sol.int1
M1: A*B + A*C + B*C -> OUT
        inclS   PRI   covS   covU
----------------------------------
1  A*B  0.956  0.931  0.375  0.179
2  A*C  0.994  0.991  0.362  0.167
3  B*C  0.994  0.991  0.434  0.239
----------------------------------
```

```
    M1  0.978  0.971  0.781
```

As the name implies, the intermediate solution sits between the conservative and parsimonious solutions in complexity: it entails three paths, as in the conservative solution, but each path now consists of only two conditions rather than three (A*B, A*C, and B*C).

We can check which logical remainder rows were used as simplifying assumptions for this solution:

```
sol.int1$SA
$M1
  A B C
8 1 1 1
```

Only row 8 was used, since row 3 was explicitly excluded from consideration.

Alternatively, we can formulate directional expectations (dir.exp) to derive an intermediate solution, rather than excluding rows by number. For example, we may expect that the presence of A and C leads to the outcome:

```
sol.int2 <- minimize(TT, include = "?", details = TRUE, use.tilde = TRUE,
                     dir.exp = "A, C")
sol.int2
From C1P1:
M1:    A*B + A*C + B*C -> OUT
        inclS   PRI   covS   covU
---------------------------------
1  A*B  0.956  0.931  0.375  0.179
2  A*C  0.994  0.991  0.362  0.167
3  B*C  0.994  0.991  0.434  0.239
---------------------------------
    M1  0.978  0.971  0.781
```

This solution is identical to sol.int1 above. When directional expectations are used, the simplifying assumptions need to be checked differently, through the i.sol component, which combines the conservative and parsimonious solutions (C1P1) and further distinguishes easy counterfactuals (EC), difficult counterfactuals (DC), and non-simplifying easy counterfactuals (NSEC). For this solution, the relevant check is:

```
sol.int2$i.sol$C1P1$EC
  A B C
8 1 1 1
```

This confirms that row 8 was treated as an easy counterfactual and used as a simplifying assumption, matching what we found through the exclude approach above.

Conjunctural directional expectations can also be specified. For example, we may expect the combination A*C, together with B on its own, to lead to the outcome:

```
sol.int3 <- minimize(TT, include = "?", details = TRUE, use.tilde = TRUE,
```

```
                    dir.exp = "A*C, B")
sol.int3
From C1P1:
M1:    B + A*C -> OUT
        inclS   PRI   covS   covU
---------------------------------
1    B  0.955  0.937  0.639  0.443
2  A*C  0.994  0.991  0.362  0.167
---------------------------------
    M1  0.964  0.953  0.805
```

This solution happens to be identical to the parsimonious solution above.

```
sol.int3$i.sol$C1P1$EC
  A B C
3 0 1 0
8 1 1 1
```

Here, both logical remainder rows 3 and 8 were treated as easy counterfactuals, which is why this set of directional expectations reproduces the parsimonious solution exactly.

As we can see, there are many different ways to derive solution terms. Beyond the standard analysis illustrated here, there are also what Schneider and Wagemann (2013) introduced as enhanced standard analysis and theory-guided enhanced standard analysis, which differ in how they treat logical remainders (see Mello 2021, Chapter 7). What is essential, regardless of approach, is that logical remainder rows are always treated in a conscious and transparent manner, with the simplifying assumptions checked and reported alongside the solution itself.

9 Visualizing Results

When applying fuzzy-set QCA, it is recommended to construct an *XY plot* that displays set-membership in the solution term against set-membership in the outcome, using the basic *plot* function of *R* or the *SetMethods* package's *xy.plot*, which adds integrated measures of fit. However, *xy.plot*'s default axes extend beyond 0 and 1 and its plots are prone to overplotting labels, so this manual instead builds plots from the ground up with *ggplot2* (Wickham 2016). Customizing *ggplot2* takes more script upfront, but the payoff is a template you can reuse across projects with minimal changes. The following sections provide such a script, show how to combine several plots on one page with *gridExtra*, and introduce the ready-made *pimplot* function from *SetMethods* for quick XY plots. Before working with *ggplot2*, it's worth browsing its documentation: https://ggplot2.tidyverse.org/

9.1 Preparation

Before creating any plots, we need a column in QCA.data that captures membership in the solution term, which will form the x-axis of the XY plot. Membership in the outcome, the y-axis, is already the OUT column we have used throughout this manual. The solution term is

calculated with the *compute* function introduced above, using the intermediate solution *sol.int1* derived in the previous chapter (A*B + A*C + B*C):

```
QCA.data$SOL <- compute("A*B + A*C + B*C", data = QCA.data)
```

We also add a column with explicit case names, which will be used further below to label individual cases in the plots:

```
QCA.data$case_name <- row.names(QCA.data)
head(QCA.data[, c("OUT", "SOL", "case_name")])
```

This confirms that both columns were added correctly, with case names ready to label individual cases in the plots that follow.

9.2 Basic XY Plot

Then, if we haven't yet done so, we install and load the *ggplot2* package (Wickham 2016):

```
install.packages("ggplot2")
library("ggplot2")
```

When creating a plot with *ggplot2*, we first specify a target object (*our.plot*), followed by the *ggplot* function and information about the data source (*QCA.data*) and the aesthetics (*aes*) to be plotted as x and y, in our case the columns *SOL* and *OUT*. Next, we indicate that the *geom_point* function creates a scatterplot. We set the point fill color (here *dodgerblue*), their size, shape, and outline color (*black*), and the line stroke (thickness). *theme_classic* gives a clean layout without gridlines, and the *theme* call that follows draws a border around the plot. *scale_x_continuous* and *scale_y_continuous* label the axes with ticks every 0.1. *coord_cartesian* extends the plotting limits slightly past 0 and 1 so corner cases stay visible. Finally, *geom_abline* adds the diagonal reference line for set-relations:

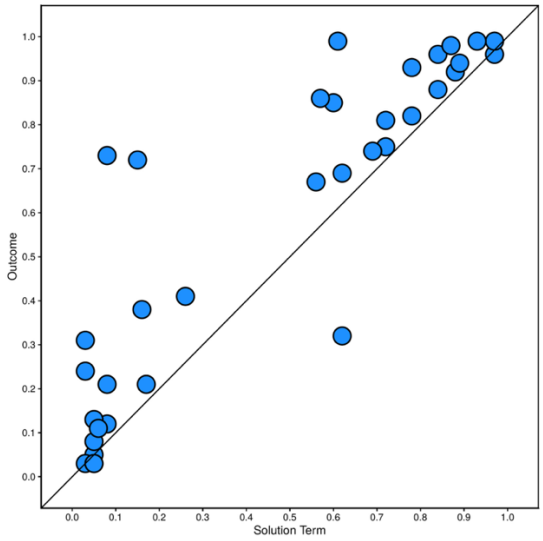
```
our.plot <- ggplot(QCA.data, aes(x = SOL, y = OUT)) +
  geom_point(
    fill = "dodgerblue",
    size = 7,
    shape = 21,
    color = "black",
    stroke = 1
  ) +
  theme_classic() +
  theme(panel.border = element_rect(
    fill = NA,
    color = "black",
    linewidth = 1,
    linetype = 1
  )) +
  scale_x_continuous(name = "Solution Term",
                     breaks = seq(0.0, 1.0, 0.1)) +
```

```
  scale_y_continuous(name = "Outcome",
                     breaks = seq(0.0, 1.0, 0.1)) +
  coord_cartesian(xlim = c(-0.02, 1.02),
                  ylim = c(-0.02, 1.02)) +
  geom_abline(intercept = 0,
              slope = 1)
our.plot
```

This script yields a basic XY plot with *ggplot2*, the result of which is shown in *Illustration 9* below. We can further use the *ggsave* function to save the plot as a PDF file (*xyplot.pdf*) with specified measurements (*width*, *height*) and resolution (*dpi*):

```
ggsave(
  "xyplot.pdf",
  plot = our.plot,
  path = NULL,
  scale = 1,
  width = 7,
  height = 7,
  dpi = 300
)
```

Illustration 9: Basic XY Plot with *ggplot2*



9.3 Refined XY Plot (Template)

The XY plot of *Illustration 9* is fine, yet it can be improved: larger text, lines marking qualitative differences, labels for typical, unaccounted, deviant, and irrelevant cases, and shading for the upper-right triangle where cases are jointly consistent with the solution term and outcome (see Mello 2021: 194). The following script achieves this and can serve as a template (also in the *RManualforQCA.R* file). First, we create subsets for typical, unaccounted, deviant, and irrelevant cases, and construct the shaded triangle's polygon shape:

```
Typ <- subset(QCA.data,
              OUT > 0.5 & SOL > 0.5)
Unacc <- subset(QCA.data,
                OUT > 0.5 & SOL < 0.5)
Devia <- subset(QCA.data,
                OUT < 0.5 & SOL > 0.5)
Irr <- subset(QCA.data,
              OUT < 0.5 & SOL < 0.5)

poly <- data.frame(x = c(0.5, 0.5, 1.1),
                   y = c(0.5, 1.1, 1.1))
```

The plot itself follows the same logic as *our.plot* above, with a few additions: the shaded triangle (*geom_polygon*), horizontal and vertical reference lines at 0.5 (*geom_hline*, *geom_vline*), larger axis text, and four calls to *geom_text_repel*, one for each subset of cases, which label the cases while avoiding overlapping text:

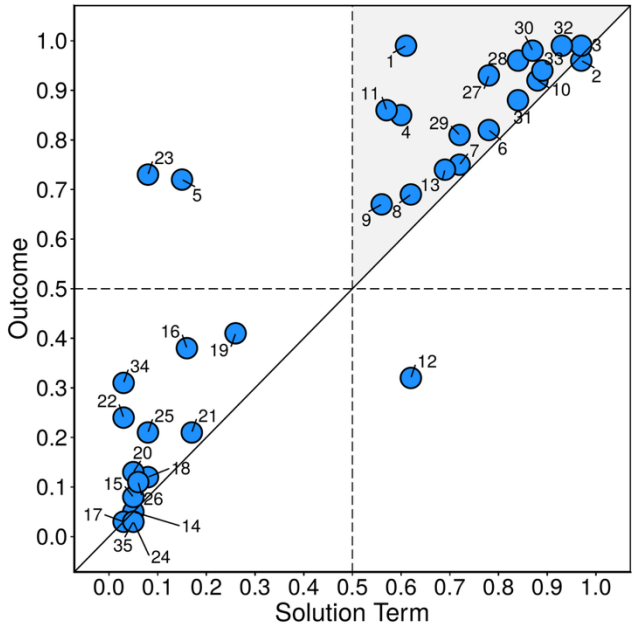
```
our.plot2 <- ggplot(QCA.data, aes(x = SOL, y = OUT)) +
  geom_polygon(
    inherit.aes = FALSE,
    aes(x = x, y = y),
    data = poly,
    fill = "gray95"
  ) +
  geom_point(
    fill = "dodgerblue",
    size = 7,
    shape = 21,
    color = "black",
    stroke = 1
  ) +
  theme_classic() +
  theme(
    panel.border = element_rect(
      fill = NA,
      color = "black",
      linewidth = 1,
      linetype = 1
    ),
    axis.text = element_text(size = 18, color = "black"),
    axis.title = element_text(face = "plain", size = 20)
  ) +
  scale_x_continuous(name = "Solution Term",
                     breaks = seq(0.0, 1.0, 0.1)) +
  scale_y_continuous(name = "Outcome",
                     breaks = seq(0.0, 1.0, 0.1)) +
```

```
  coord_cartesian(xlim = c(-0.02, 1.02),
                  ylim = c(-0.02, 1.02)) +
  geom_hline(yintercept = 0.5,
             color = "black",
             linetype = 5) +
  geom_vline(xintercept = 0.5,
             color = "black",
             linetype = 5) +
  geom_abline(intercept = 0,
              slope = 1) +
  geom_text_repel(data = Unacc,
                  aes(label = case_name),
                  box.padding = 0.5,
                  size = 5,
                  max.overlaps = 20) +
  geom_text_repel(data = Devia,
                  aes(label = case_name),
                  box.padding = 0.5,
                  size = 5,
                  max.overlaps = 20) +
  geom_text_repel(data = Typ,
                  aes(label = case_name),
                  box.padding = 0.5,
                  size = 5,
                  max.overlaps = 20) +
  geom_text_repel(data = Irr,
                  aes(label = case_name),
                  box.padding = 0.5,
                  size = 5,
                  max.overlaps = 20)
our.plot2
```

This produces the XY plot shown in *Illustration 10* below. We save it the same way as before:

```
ggsave(
  "xyplot2.pdf",
  plot = our.plot2,
  path = NULL,
  scale = 1,
  width = 7,
  height = 7,
  dpi = 300
)
```

Illustration 10: Enhanced XY Plot with *ggplot2*



In case we want to visualize the constituent paths of a QCA solution, we can combine several plots on the same page layout, as shown in the next two sections. The argument *geom_text_repel* serves to avoid overplotting labels; this too can be customized, for instance if you prefer not to have lines indicating the positions of the cases.

9.4 Combining XY Plots (Template)

Each of the three sufficient paths of *sol.int1* (A*B, A*C, and B*C, see the chapter on solution types) can be plotted the same way as the overall solution term above, showing which cases are typical, deviant, unaccounted for, or irrelevant with respect to that specific path. Rather than repeating the same block of *ggplot2* code three more times, we wrap it in a function and call it once per path:

```
make.xyplot <- function(data, xvar, xlab) {

  data$X <- data[[xvar]]

  Typ   <- subset(data, OUT > 0.5 & X > 0.5)
  Unacc <- subset(data, OUT > 0.5 & X < 0.5)
  Devia <- subset(data, OUT < 0.5 & X > 0.5)
  Irr   <- subset(data, OUT < 0.5 & X < 0.5)

  poly <- data.frame(x = c(0.5, 0.5, 1.1),
                     y = c(0.5, 1.1, 1.1))

  ggplot(data, aes(x = X, y = OUT)) +
    geom_polygon(
      inherit.aes = FALSE,
```

```
    aes(x = x, y = y),
    data = poly,
    fill = "gray95"
  ) +
  geom_point(
    fill = "dodgerblue",
    size = 7,
    shape = 21,
    color = "black",
    stroke = 1
  ) +
  theme_classic() +
  theme(
    panel.border = element_rect(
      fill = NA,
      color = "black",
      linewidth = 1,
      linetype = 1
    ),
    axis.text = element_text(size = 18, color = "black"),
    axis.title = element_text(face = "plain", size = 20)
  ) +
  scale_x_continuous(name = xlab,
                     breaks = seq(0.0, 1.0, 0.1)) +
  scale_y_continuous(name = "Outcome",
                     breaks = seq(0.0, 1.0, 0.1)) +
  coord_cartesian(xlim = c(-0.02, 1.02),
                  ylim = c(-0.02, 1.02)) +
  geom_hline(yintercept = 0.5,
             color = "black",
             linetype = 5) +
  geom_vline(xintercept = 0.5,
             color = "black",
             linetype = 5) +
  geom_abline(intercept = 0,
              slope = 1) +
  geom_text_repel(data = Unacc,
                  aes(label = case_name),
                  box.padding = 0.5,
                  size = 5,
                  max.overlaps = 20) +
  geom_text_repel(data = Devia,
                  aes(label = case_name),
                  box.padding = 0.5,
                  size = 5,
```

```
                    max.overlaps = 20) +
    geom_text_repel(data = Typ,
                    aes(label = case_name),
                    box.padding = 0.5,
                    size = 5,
                    max.overlaps = 20) +
    geom_text_repel(data = Irr,
                    aes(label = case_name),
                    box.padding = 0.5,
                    size = 5,
                    max.overlaps = 20)
}
```

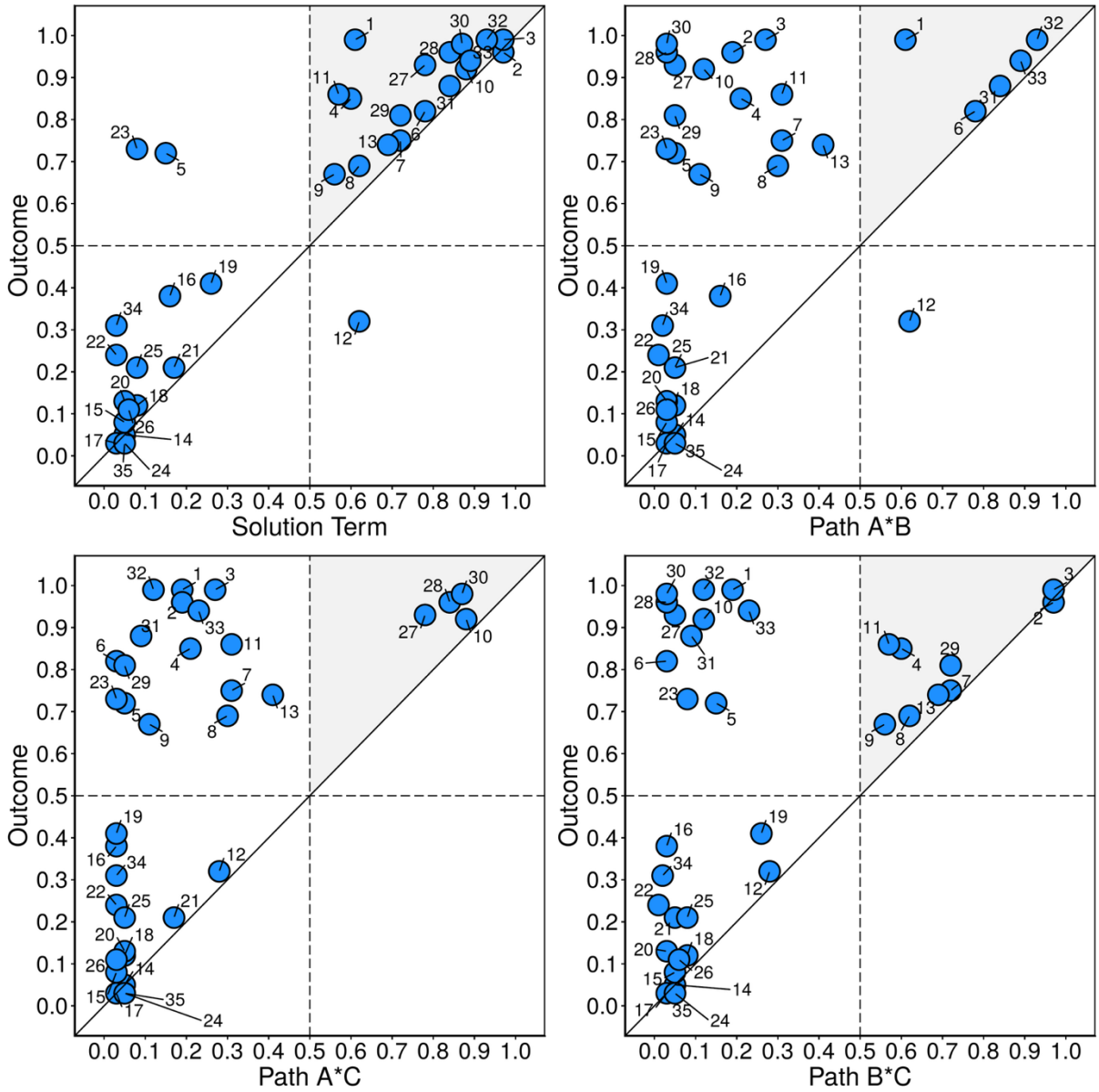
We then compute membership in each individual path and build one fully customized plot per path (*our.plot2*, the plot for the overall solution term, was already created above):

```
QCA.data$P1 <- compute("A*B", data = QCA.data)
QCA.data$P2 <- compute("A*C", data = QCA.data)
QCA.data$P3 <- compute("B*C", data = QCA.data)

plot.P1 <- make.xyplot(QCA.data, "P1", "Path A*B")
plot.P2 <- make.xyplot(QCA.data, "P2", "Path A*C")
plot.P3 <- make.xyplot(QCA.data, "P3", "Path B*C")
```

Finally, we combine the overall-solution plot and the three path plots on a single PDF page, arranged in two rows and two columns. This requires the *gridExtra* package and its *grid.arrange* function; note that the `par(mfrow = ...)` approach used for the histograms above only affects base *R* graphics and has no effect on *ggplot2* plots such as these:

```
install.packages("gridExtra")
library("gridExtra")

pdf(file = "xyplot_combined.pdf", width = 12, height = 12)
grid.arrange(our.plot2, plot.P1, plot.P2, plot.P3, ncol = 2, nrow = 2)
dev.off()
```

Illustration 11: Combined XY Plots with *gridExtra*



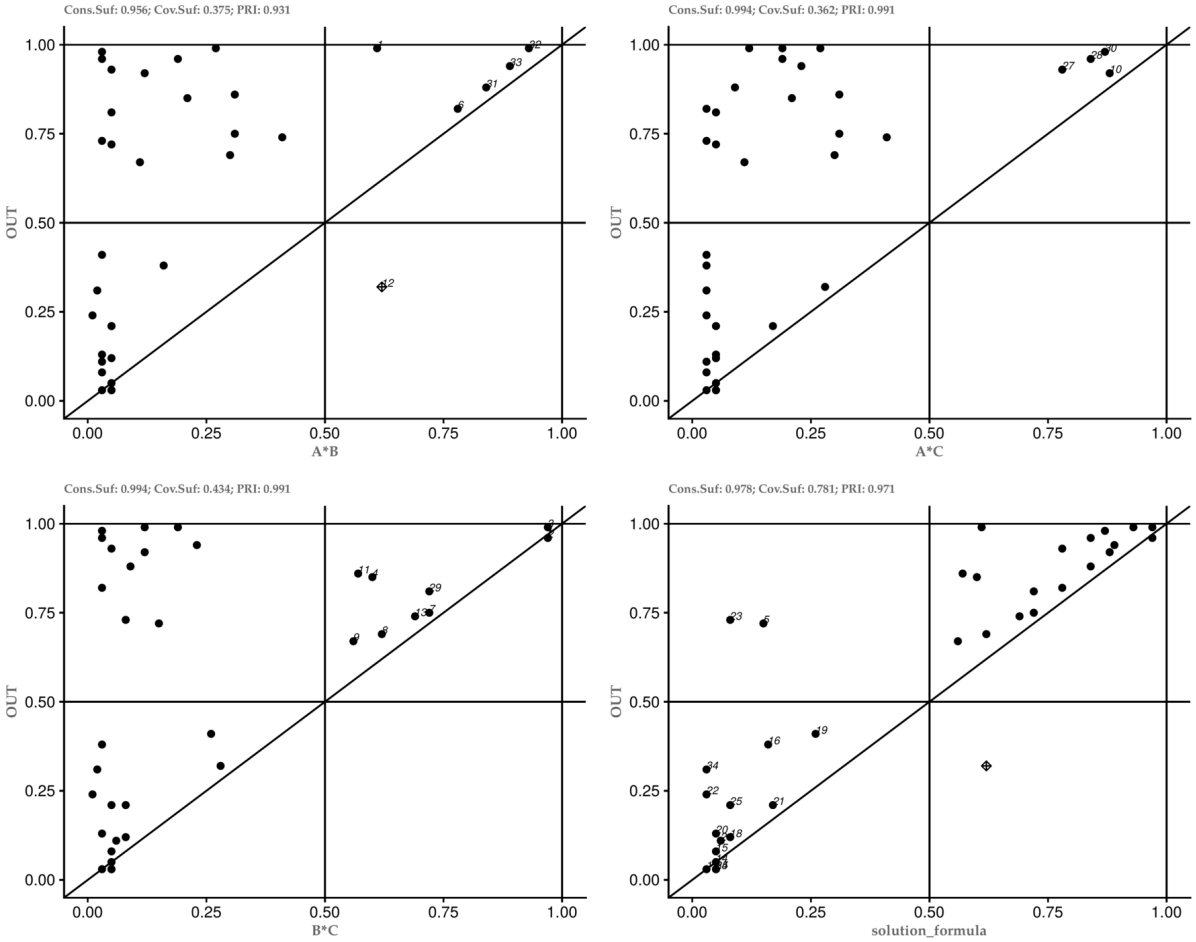
9.5 Pimplot Function

For a quicker alternative, the *pimplot* function from the *SetMethods* package automatically creates XY plots for each sufficient path along with the overall solution, without requiring a custom script:

```
pimplot(data = QCA.data,
        results = sol.int1,
        outcome = "OUT")
```

Unlike the approach above, *pimplot* prints each of these plots separately rather than combining them on a single page, and it offers less customization: no labeled subsets, polygon shading, or threshold lines. For compactness, *Illustration 12* below arranges the four resulting plots (the overall solution term, and its three constituent paths A*B, A*C, and B*C) together.

Illustration 12: XY Plots with *pimplot*



This is just one example of how QCA results can be visualized and how the *ggplot2* package (Wickham 2016) can be used for that purpose. I also recommend having a look at visualization options explored by Rubinson (2019) and Oana, Schneider & Thomann (2021), among others.

10 Cluster Analysis

The *cluster* function of the *SetMethods* package extends the QCA framework to clustered data, where cases are grouped into higher-level units such as regions or time periods. This lets us assess whether solution terms derived from the pooled analysis hold consistently across clusters, or whether they mask meaningful differences between groups.

Following Oana, Schneider & Thomann (2021), we illustrate this using the example of the study by Paykani et al. (2018) entailed in the PAYF data set from the *SetMethods* package, a clustered example in which cases (countries) are nested within regions:

```
data(PAYF)
TTpayf <- truthTable(PAYF, "HL", conditions = c("HE", "GG", "AH", "HI",
"HW"), complete = TRUE, show.cases = TRUE, incl.cut = 0.75, sort.by =
"incl, n")
TTpayf
```

```
 OUT: output value
   n: number of cases in configuration
incl: sufficiency inclusion score
 PRI: proportional reduction in inconsistency

    HE GG AH HI HW   OUT    n   incl  PRI
30   1  1  1  0  1     1     36  0.929 0.906
32   1  1  1  1  1     1     5   0.925 0.824
24   1  0  1  1  1     1     2   0.890 0.470
20   1  0  0  1  1     1     2   0.873 0.281
14   0  1  1  0  1     1     1   0.868 0.435
26   1  1  0  0  1     1     1   0.855 0.375
17   1  0  0  0  0     0     1   0.747 0.110
 3   0  0  0  1  0     0     30  0.332 0.036
```

(This table has been shortened to the top eight truth table rows, sorted by consistency)

The country codes for each configuration are likewise omitted; running the command with *show.cases = TRUE* reproduces the full listing. The configuration in row 30 (HE, GG, AH, and HW present, HI absent, n = 36), groups most high-income countries, from Australia and Austria to the United Kingdom and the United States. Conversely, row 3 at the bottom of the excerpt above, is dominated by lower- and middle-income countries such as Angola, Bangladesh, and Kenya (only HI present, n = 30, coded 0 on the outcome).

From the truth table, we derive the conservative solution:

```
SOLpayf <- minimize(TTpayf, include = "1", details = TRUE, use.tilde =
TRUE)
SOLpayf
M1: HE*HI*HW + AH*~HI*HW + ~HE*GG*AH*~HW + HE*~GG*~AH*HI + HE*~GG*AH*~HI
+ HE*GG*~AH*~HI + ~GG*~AH*HI*HW -> HL
                  inclS   PRI   covS   covU
------------------------------------------
1       HE*HI*HW  0.862  0.688  0.334  0.075
2      AH*~HI*HW  0.863  0.820  0.638  0.435
3  ~HE*GG*AH*~HW  0.803  0.225  0.149  0.004
4  HE*~GG*~AH*HI  0.804  0.216  0.200  0.005
5  HE*~GG*AH*~HI  0.702  0.160  0.141  0.001
6  HE*GG*~AH*~HI  0.788  0.274  0.143  0.004
7  ~GG*~AH*HI*HW  0.760  0.208  0.213  0.019
------------------------------------------
             M1  0.789  0.702  0.862
```

As with the truth table above, the case listing under each path has been omitted here for reasons of space. Path 2 (AH*~HI*HW), which has by far the highest unique coverage (covU = 0.435), covers the large group of high-income countries identified in the truth table, plus Azerbaijan, Serbia, Kazakhstan, and Russia.

The *cluster* function tests this solution against the cluster structure, here defined by country (unit_id) nested within region (cluster_id):

```
cluster.sol <- cluster(results = SOLpayf, data = PAYF, outcome = "HL",
unit_id = "COUNTRY", cluster_id = "REGION", wicons = FALSE)
cluster.sol
```

The full *cluster.sol* output reports consistencies, distances, and coverages for all seven paths across the pooled sample and each of the six regions included in the study. To keep this manageable, the tables below have been shortened to two of the seven paths, chosen because they show the clearest contrast between regions. The distances information is given in full, as it condenses to a single value per path.

```
Consistencies (2 of 7 paths shown):
                                        HE*HI*HW  ~GG*~AH*HI*HW
Pooled                                     0.862          0.760
Between African Region (33)                0.279          0.152
Between Eastern Mediterranean Region (8)   0.916          0.900
Between European Region (46)               1.000          1.000
Between Region of the Americas (26)        0.905          0.815
Between South-East Asian Region (8)        0.798          0.822
Between Western Pacific Region (10)        0.744          0.878

Coverages (2 of 7 paths shown):
                                        HE*HI*HW  ~GG*~AH*HI*HW
Pooled                                     0.334          0.213
Between African Region (33)                0.522          0.391
Between Eastern Mediterranean Region (8)   0.782          0.790
Between European Region (46)               0.175          0.072
Between Region of the Americas (26)        0.610          0.403
Between South-East Asian Region (8)        0.691          0.711
Between Western Pacific Region (10)        0.295          0.175

Distances (From Between to Pooled, all 7 paths):
HE*HI*HW        0.125
AH*~HI*HW       0.097
~HE*GG*AH*~HW   0.147
HE*~GG*~AH*HI   0.148
HE*~GG*AH*~HI   0.130
HE*GG*~AH*~HI   0.099
~GG*~AH*HI*HW   0.150
```

There is a clear pattern: consistency for HE*HI*HW and ~GG*~AH*HI*HW both hold up well in the European region (1.000 in each case) but drop sharply in Africa (0.279 and 0.152). Coverage tells the opposite story: HE*HI*HW covers only 0.175 of European cases, but covers 0.782 in the Eastern Mediterranean. Distances confirm every path drifts from the pooled score

once broken down by region, from 0.097 (AH*~HI*HW, the most stable) to 0.150 (~GG*~AH*HI*HW, the least stable), which illustrates the heterogeneity the *cluster* function is designed to identify.

11 Robustness Tests

11.1 Preparation

Robustness tests serve to "establish the extent to which the results of the set-theoretic analysis are conditional upon specific analytical choices" (Mello 2021, 208). This can involve broader research design decisions, which would call for qualitative changes. In a narrower sense, robustness tests are used to assess the sensitivity of QCA results to analytical decisions such as calibration anchors or the consistency threshold.

We can conduct such tests with the protocol designed by Oana and Schneider (2021), implemented in *SetMethods*. These tests require both raw and calibrated data, so instead of the QCAdata.csv file used previously, we use data from a published study: Meissner and Mello (2022) on the unintended consequences of UN sanctions (SanctionsData.csv). The study and its replication data are open access, and the article's online appendix documents the robustness tests in detail. The original data has been shortened here for illustration.

Article and appendix: https://doi.org/10.1080/13523260.2022.2059226
Data: https://doi.org/10.7910/DVN/KYF7VJ

```
SD <- read.csv("SanctionsData.csv", row.names = 1)
```

We separate the raw and calibrated columns into two data frames, as required by the robustness functions, then create a truth table and derive the initial solution, which replicates the intermediate solution reported in Meissner and Mello (2022):

```
SD.raw   <- SD[, 1:6]
SD.calib <- SD[, 7:12]
TT.s <- truthTable(SD.calib, "UC", conditions = c("AR", "LS", "P5", "EI",
"LD"), complete = TRUE, show.cases = TRUE, incl.cut = 0.75, sort.by =
"incl, n")
IS <- minimize(TT.s, include = "?", details = TRUE, use.tilde = TRUE,
exclude = c(1, 5, 9, 17))
```

The two configurations with the highest inclusion scores both show the outcome: LS and P5 present with EI absent (row 14, incl = 1.000, Côte d'Ivoire and Somalia) and the same configuration with LD also absent (row 13, incl = 0.980, Libya 2). At the other end, closest to consistently negative are AR and EI present with P5 absent (row 19, incl = 0.091, Sudan 1) and AR, EI, and LS all present (row 23, incl = 0.151, Iran).

```
IS
M1: LS*P5 + ~AR*~P5*LD + AR*P5*~EI -> UC
```

```
               inclS   PRI   covS   covU
-----------------------------------------
1       LS*P5  0.935  0.931  0.506  0.251
   cases: Libya 2; Côte d'Ivoire,Somalia; DPRK; Angola; Haiti
2  ~AR*~P5*LD  0.936  0.935  0.231  0.229
   cases: DRC; Liberia; Sierra Leone
3   AR*P5*~EI  0.910  0.902  0.366  0.111
   cases: FRY; Libya 1; DPRK; Angola
-----------------------------------------
          M1  0.922  0.919  0.847
```

The initial solution (IS) combines three sufficient paths to withdrawal of UN sanctions (UC), each with consistency (inclS) above 0.90. The overall solution reaches a consistency of 0.922 and covers 0.847 of cases with the outcome.

11.2 Calibration Sensitivity

The *rob.calibrange* function identifies the range of calibration anchors for a given condition within which the solution formula does not change. Wide ranges indicate that a calibration choice is robust, while narrow ranges signal that the result depends heavily on where the anchors were placed. Here we test the sensitivity of the outcome's anchors (full exclusion at 0.5, cross-over at 0.75, full inclusion at 1.25, per Meissner & Mello 2022, Table 1):

```
rob.calibrange(raw.data = SD, calib.data = SD.calib,
               test.cond.raw = "UC_raw", test.cond.calib = "UC",
               test.thresholds = c(0.5, 0.75, 1.25),
               step = 0.01, max.runs = 100,
               outcome = "UC",
               conditions = c("AR", "EI", "LS", "LD", "P5"),
               incl.cut = 0.75, n.cut = 1, include = "?",
               exclude = c(1, 5, 9, 17))

Exclusion:  Lower bound  NA Threshold  0.5 Upper bound  0.5
Crossover:  Lower bound  0.51 Threshold  0.75 Upper bound  1.24
Inclusion:  Lower bound  0.76 Threshold  1.25 Upper bound  NA
```

These results indicate considerable room to move the calibration anchors without changing the sufficient solution. The cross-over point shows the widest margin, from 0.51 to 1.24 around its original 0.75. The inclusion anchor could be lowered to 0.76 before the solution would change, with no upper bound established. The exclusion anchor's upper bound coincides with its original threshold of 0.5, though its lower bound is likewise unconstrained. Overall, the calibration choices reported in Meissner and Mello (2022, Table 1) appear robust, particularly around the cross-over point.

11.3 Test Solution Robustness

Beyond calibration, we can also compare the initial solution against alternative test solutions derived under different analytical choices, such as a higher consistency threshold or a different calibration for the outcome. The example below constructs two such test solutions, then evaluates them jointly with *rob.corefit* (which reports fit for the robust core, meaning the overlap between the initial and all test solutions) and *rob.fit* (which reports overall robustness parameters, where values close to 1 for RFcons and RFcov indicate high robustness):

```
TT.t1 <- truthTable(SD.calib, "UC",
                    conditions = c("AR", "LS", "P5", "EI", "LD"),
                    complete = TRUE, show.cases = TRUE,
                    incl.cut = 0.80, sort.by = "incl, n")
TS1 <- minimize(TT.t1, include = "?", details = TRUE,
                use.tilde = TRUE, exclude = c(1, 5, 9, 17))

SD.calib2 <- SD.calib
SD.calib2$UC <- calibrate(SD.raw$UC_raw, type = "fuzzy",
                          method = "direct", thresholds = c(0, 1, 2))

TT.t2 <- truthTable(SD.calib2, "UC",
                    conditions = c("AR", "LS", "P5", "EI", "LD"),
                    complete = TRUE, show.cases = TRUE,
                    incl.cut = 0.75, sort.by = "incl, n")
TS2 <- minimize(TT.t2, include = "?", details = TRUE,
                use.tilde = TRUE, exclude = c(1, 5, 9, 17))
TS <- list(TS1, TS2)

rob.corefit(test_sol = TS, initial_sol = IS, outcome = "UC")
Cons.Suf  Cov.Suf      PRI
   0.935    0.736    0.932

rob.fit(test_sol = TS, initial_sol = IS, outcome = "UC")
               RF_cov RF_cons RF_SC_minTS RF_SC_maxTS
Robustness_Fit  0.869   0.986       0.857          1
```

The *rob.corefit* function compares the initial solution (IS) against the test solutions (TS1 and TS2, combined into TS). The scores are high but still less than 1, which means that there is no perfect overlap between the robust core and the initial solution, which is to be expected based on the way the robust core is calculated (Oana & Schneider 2024, 77). The robustness fit scores (RF_cov = 0.869, RF_cons = 0.986) point the same way, and the two bounds (RF_SC_minTS = 0.857, RF_SC_maxTS = 1) stay close to the maximum, indicating that the sufficient solution holds up well against the analytical changes applied in the test sets. For further details on the calculation and interpretation of these robustness measures, see Oana and Schneider (2024).

12 References

Duşa, A. 2019. *QCA with R. A Comprehensive Resource*. Cham: Springer.

Duşa, A. 2022. *Venn: Draw Venn Diagrams*. R Package Version 1.11.

Elff, M. 2021. *Data Management in R*. London: Sage.

Field, A., Miles, J., & Field, Z. 2012. *Discovering Statistics Using R*. Los Angeles: Sage.

Gaubatz, K. T. 2015. *A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved*. Los Angeles, CA: Sage.

Imai, K. 2017. *Quantitative Social Science: An Introduction*. Princeton, NJ: Princeton University Press.

Kabacoff, R. I. 2015. *R In Action: Data Analysis and Graphics with R*. Shelter Island: Manning.

Meissner, K. L., & Mello, P. A. 2022. The Unintended Consequences of UN Sanctions: A Qualitative Comparative Analysis. *Contemporary Security Policy* 43 (2): 243-73.

Mello, P. A. 2021. *Qualitative Comparative Analysis: An Introduction to Research Design and Application*. Washington, DC: Georgetown University Press.

Oana, I.-E., & Schneider, C. Q. 2018. SetMethods: An Add-on R Package for Advanced QCA. *The R Journal* 10 (1): 507-33.

Oana, I.-E., & Schneider, C. Q. 2024. A Robustness Test Protocol for Applied QCA: Theory and R Software Application. *Sociological Methods and Research* 53 (1): 57-88.

Oana, I.-E., Schneider, C. Q., & Thomann, E. 2021. *Qualitative Comparative Analysis Using R: A Beginner's Guide*. New York: Cambridge University Press.

Paykani, T., Rafiey, H., & Sajjadi, H. 2018. A Fuzzy Set Qualitative Comparative Analysis of 131 Countries: Which Configuration of the Structural Conditions Can Explain Health Better?. *International Journal for Equity in Health*, 17 (1).

Pollock, P. H., & Edwards, B. C. 2018. *An R Companion to Political Analysis*. Thousand Oaks: Sage.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*, Vienna, R Foundation for Statistical Computing.

RStudio (2020) *RStudio: Integrated Development Environment for R*, Boston.

Rubinson, C. 2019. Presenting Qualitative Comparative Analysis: Notation, Tabular Layout, and Visualization. *Methodological Innovations* 12 (2): 1-22.

Schneider, C. Q., & Wagemann, C. 2013. Doing Justice to Logical Remainders in QCA: Moving Beyond the Standard Analysis. *Political Research Quarterly* 66 (1): 211-20.

Vis, B. 2009. Governments and Unpopular Social Policy Reform: Biting the Bullet or Steering Clear? *European Journal of Political Research* 48 31–57.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.