RELAN - Relation Analysis

RELAN User Manual - English

This guide is intended to provide interested laymen and empirical practitioners with a rough overview of how to call up the various applications of the program, which inputs are necessary for which problems and which outputs can be expected for the various options. With the available sample files and the free software of the program (RELAN Beta 1.0), tests of RELAN can be carried out immediately. For detailed theoretical justifications or methodological explanations, the reference literature must be consulted:
Maderthaner, R. (2022). Relationsanalyse (RELAN) - Aussagenlogische, statistische und kausale Analyse von Daten.

For a test run of the program (RELAN Program), only four files are immediately important, namely RELAN.exe, RELAN-IN, RELAN-OUT and possibly RELAN-DAT, which are available together with other input examples in the directory Probe-Files (Testing Files). The quickest way to familiarize with the program is to start the free software RELAN_Beta 1.0, so that it runs with the available input file (RELAN-IN) and one can get an impression of the resulting output (RELAN-OUT) by the additional use of the OUTPUT chapter of this manual. If extended experiences are desired, the input parameters can be easily modified directly by changing RELAN-IN or by using the Graphical User Interface (GUI) or by copying the Input blocks of the output examples (Testing Files) for further runs.

A total of ten files can be used in the program for the different options, which are characterized in more detail in the RELAN Fact Sheet (RELAN-IN, RELAN-FCT, RELAN-WIE, RELAN-DAT, RELAN-KAT, RELAN-OUT, RELAN-EXP, RELAN-GRA, RELAN-TRU). The program should run in its own directory, because all mentioned files are automatically created (or overwritten) there.

OBJECTIVE

- Natural **language hypotheses** can be statistically tested on a propositional logical basis by linking their linguistic components with

logical connections (AND, OR, IF-THEN, ...), whereby many relationships between variables, which have so far hardly been considered in statistics, can be analysed. For example, the hypothesis that vaccination (VAC) AND antibody formation (ANT) OR recovery of COVID disease (REC) cause protection (PRO) against COVID disease can be formulated logically and statistically tested: ((VAC * ANT) + REC) > PRO (see also RELAN Examples).

- In one hypothesis, different **types of influences** can be considered by differentiating the set variables (cause, effect, moderator, mediator, and condition).
- For the evaluations, different **chance models** can be selected for computing the expected values of the variables, so that results from different investigations can be compared more fairly.
- It is possible to subtract **falsifications** (confounding variables or groups of confounding variables) from the effects of hypotheses.
- **Causal analyses** can be carried out in which the logical and temporal prerequisites required for causal interpretation are actually given (IF-THEN relations between variables, temporal coincidence of causes and effects).
- In addition, the program enables comprehensive exploratory analyses ("data mining") by taking into consideration all existing logical regularities (relations) in a given set of (bivalent) variables.

INPUT

The input file RELAN-IN is a text file in which all entries can be made directly or - more convenient and clear - using the so-called "Graphical User Interface" (GUI). In order to get a quick impression of the program's capabilities, it is also possible to use the input blocks of the available output files (see RELAN Examples, Testing Files), which are available on the first page of each RELAN-OUT file and can be copied into a RELAN-IN.

Important: For each type of evaluation (options), all parameters must always be given in RELAN-IN, even those that are not required in the currently selected evaluation run.

The annotated RELAN-IN file in PDF format (RELAN-IN-Commentary) can also serve as a short instruction, in which some explanations are

attached for each parameter. In the input file each parameter type (title, option, input, number of variables, ...) is preceded by a line in which the name and meaning of the respective parameter is noted and this line also can be used for individual comments, because it is not considered in the program.

For a better understanding of the following listing of the parameters, the sample file RELAN-IN should also be opened in the meantime.

TITEL: Name of the evaluation (no more than 80 letters)

OPTION: By this parameter the type of desired evaluation is decided: SIMULATION = 3: Simulation and statistical evaluation of propositionally formulated hypotheses (relations), which contain the names of the variables (e.g., REC, VAC, ANT, ...) as well as propositional logical links, such as * (AND), + (OR), > (IF-THEN), = IF and only IF-THEN, -* (Not AND). (see also RELAN Examples)

CONFIRMATION = 2: Statistical testing of propositional hypotheses with output of descriptive statistics, significance values and effect sizes (see also RELAN Statistics). The hypothetical relation is evaluated on the basis on about 30 statistics and can also be compared with other relations with regard to significance and generalizability.

IMPLICATIONS = 1: Exploration of implications between all pairs of variables to check whether there are left-sided or right-sided IF-THEN relations between them. More than 30 common statistics are computed for all significant relations (see also RELAN Statistics). In addition, a graph-theoretical analysis of the significant implications can be requested, whereby the different networks between the variables are shown (output, switching end nodes, branches, etc.). A causal analysis can elucidate causal relationships within a network of significant variables.

CONNECTIONS = 0: Exploration of connections (seven propositional logical links) between the variables. A total of seven possible relations between two variables are checked and statistically evaluated. (see RELAN Statistics). A graph-theoretical analysis is also provided for this evaluation option.

DATA INPUT: The data can be read either in matrix format (in the file RELAN-DAT) or in vector format (in the file RELAN-KAT). In the matrix format, the data is arranged (as usual) in such a way that a case with its variable values (separated by spaces) is entered in each line (see, for example, RELAN-DAT). In the less frequently used vector format, all associated case frequencies must be specified for all (2^M) expression combinations (Minterme) of the given (M) variable. The order of entry complies with the truth table of the variables involved (e.g. 0-0-0, 0-0-1, 0-1-0, etc.).

NUMBER OF VARIABLES (M): The only difference between the free and licensed versions is the number of possible variables and the scope of possible cases (N) that can be used for the analyses. In the free version up to 4 variables and up to 50 cases are possible, in the licensed version 100 variables and 1000 cases.

DICHOTOMISIERUNG: Here you have to specify whether the data are already dichotomized (i.e., only have values of 0 and 1) or whether cutting scores must be used for each variable in the VERCTOR OF DICHOTOMISATION. Any value in the data that is equal to or greater than the corresponding cutting score is converted to 1 (one), the remaining values are assigned 0 (zero).

NAMES OF VARIABLES: All variables have three-digit names, which must be entered separated by one or several spaces.

VEKTOR OF DICHOTOMISIERUNG: Import of the cutting scores for all variables. If the value of a variable is equal to or greater than its cutting score, it is assigned the value 1, otherwise the value 0. With planned dichotomization, it is advantageous to enter the cutting scores for the variables exactly under the variable names, so that a reliable assignment is ensured. Cutting scores can be entered with or without decimal point.

PROBABILITIES OF THE APRIORI CHANCE MODEL: For each variable, estimates of probabilities of its positive values must be entered (default value: p = 0.5). The probabilities of the variables can either be hypothetical, come from other studies or be taken from statistical databases. These are then considered as APRIORI CHANCE MODEL. If this is to be used for the calculations, then the value "apriori" must be selected

for the parameter HYPOTHESIS CHANCE MODEL. If, on the other hand, the relative frequencies of the variables from the sample are to be used as probabilities for the calculations, then "aposteriori" must be selected for the HYPOTHESIS CHANCE MODEL.

ORDER OF CAUSALITY: In causal analysis for each variable a natural number (e.g., 1, 2, 3, ...) has to be determined to symbolize the point in time at which the respected variables (interpreted as events) can have an effect on other variables.

DURATION OF CAUSALITY: In causal analyses also for each variable has to be fixed with a natural number (e.g., 2, 2, 3, ...) over how many units of time the possible effect is considered to be effective (see also RELAN Examples).

PRINT EXTENT: This parameter specifies the amount of output of calculations in RELAN-OUT (0 = minimum, 3 = maximum).

HYPOTHESIS CHANCE MODEL: Here, the desired chance model is selected (null hypothesis), according to which the significance calculations are performed. If 1 (= a priori) is selected, then the theoretically determined probabilities are used, and at 0 (= aposteriori) the relative frequencies of the variables calculated from the sample are used (for the calculation of the expected values).

LEVEL OF SIGNIFICANCE (Zsig): Determination of the probability of error (alpha error) for the significance checks. For one-sided questions (e.g., whether frequencies are greater than the expected value), this value is 1,645 at 5 percent and 2,330 at 1 percent of error probability.

LEVEL OF SIMULATION (Zsim): Determination of the desired significance level for a simulation of a propositional logical hypothesis in a fictitious sample. The larger this value is chosen, the more the data are fitted to the postulated propositional logical hypothesis. For extremely concise fittings of the data to a hypothesis, even Zsim values up to 30 can be used.

LEVEL OF EXTRACTION (Zext): Determination of the desired significance level for a (confounding) relation to be subtracted from the hypothesis function. The higher this value, the more the function to be

calculated deviates from the expected values. If 0.0 is entered here, no extraction relation is calculated.

INCLUSION CRITERION: The INCLUSION CRITERION can be non-normalized or normalized (at the options of CONNECTIONS or IMPLICATIONS). Since the minima and maxima of Z-values also depend on the number of truth values of a relation, normalized Z-values also are computed in addition to the usual Z-values. When entering 0 (not normalised), usual significance calculations are calculated for the relations, when entering 1 (normalised), the z-values are calculated as if all relations had the same number of positive and negative truth values.

Y(P)-CUT und X(P)-CUT: Since pairwise testing of variables can result in many combinations, it is possible in the program to limit the output twice by specifying a limit for the relative frequency of the first variable X and/or the second variable Y, so that only those results are printed for which the relative frequencies are above the respective limits. This also considers the practical importance of variables.

GRAPH THEORETICAL ANALYSIS: With the option's "connections" and "implications", analyses which based on graph theory can also be carried out for significant relations. This is used to identify relation chains, determine their output and end variables and show circular variable relationships. In order to be able to process the outputs in graphics programs (e.g., yEd, Decision Explorer), the with variable pairs, relation weights, and relation types are written out in the RELAN-GRA file.

WEIGHTING OF CASES: If there are doubts about the empirical credibility of some cases (poor survey conditions, uncertain measurements, misleading instructions, ...), then this can be expressed with a credibility index between 0.0 (not usable) and 1.0 (full usability). When entering data in matrix format (using RELAN-DAT), the individual cases can be weighted with index values between 0.0 and 1.0 according to their empirical credibility. These weighting values must be entered for all cases in the existing order in the RELAN-WEI file (series of numbers with spaces as space).

CAUSAL ANALYSIS: If a causal analysis is to be carried out (1 = yes), then it is assumed that the values of variables are to be interpreted as

"events" to which effects are attributed from a certain point in time and for a certain duration. For the assumption of causally interactions the values of variables (symbolic events) must be realized in the same time level. The cause event must coincide with the effect event at a certain point in time for an effect to occur. Whether this coincidence of effects can be assumed is examined in the program based on the defined impact indicators (ORDER and DURATION OF CAUSALITY). A relation is categorized as causally related if all its logical components are realized in at least one temporal period (defined by the indices).

TRUTH FUNCTION: Propositional relations can be characterized either in the form of functions (e.g. (VAC + REC) > PRO) or in the form of truth tables (e.g., 1 0 0 1 1 0 etc.). The parameter is 0 (= no) if the relation of the HYPOTHETICAL FUNCTION is entered as a propositional formula (in RELAN-IN), and 1 (= yes) if the truth function of the hypothetical relation is imported via the RELAN-FCT file (1 = true, 0 = not true). All produced truth functions developed in the evaluation are printed in the RELAN-TRU file.

HYPOTHETICAL FUNCTION: This parameter input contains a propositional logical function (e.g., VAC + REC) > PRO), which is simulated to the extent of the LEVEL OF SIMULATION or tested to the extent of the LEVEL OF SIGNIFICANCE. If several unconnected functions are entered (for propositional comparison purposes), then only the last function is statistically evaluated. The function cannot exceed 80 characters (including spaces).

EXTRACTIONAL FUNCTION: This parameter input contains a propositional logical function that is simulated to the extent of the LEVEL OF EXTRACTION and then eliminated (reduced) from the HYPOTHETICAL FUNCTION. The function formula cannot exceed 80 characters (including spaces). If only the effect of a single variable is "partialized", then this must be entered as a bivariate function (e.g. REC * REC).

DATEN

Most often, programs enter data in conventional form, namely as a matrix, with each case positioned in a row with the corresponding number of variable values. For each case, the same number of variable values

must be entered in RELAN, a replacement of missing values ("imputation") is not provided in the current program version. The variable values must be entered in free format, i.e., they can contain decimal points and must be separated by a space (or more) (see RELAN-DAT). Another, probably much less frequently chosen input type is that in the form of a series of numbers ("vector"), in which the case numbers attributable to each combination of the variables (minterms) must be entered (RELAN-KAT). The order of the input is based on the truth table for the corresponding variable set and can be found in the relevant table from each output file. Since vector input of the data is already aggregated (sum of cases per combination of characteristics), the optional case weighting must be omitted here. In the RELAN-EXP file, the data obtained by simulation or those that have been read in are exported so that they are not only processed in other programs (e.g. SPSS, SAS), but (after deleting the variable name line) can also be imported into the RELAN-DAT file for further analysis.

PROGRAM FLOW

When the program starts, three (green) windows open. The right window shows the types of files present or formed in the directory, the first left window illustrates the running output (RELAN-OUT), the second left window (behind) displays the input parameters (RELAN-IN). Without faults a box with the text "Program terminated, exit code 0 – Exit Window" appears and "yes" closes all windows. The same effect is achieved with "CTRL C", which can also be used to stop the calculations of RELAN if waiting times are too long. If the program has been terminated prematurely, a justification appears in the output file (RELAN-OUT) for most errors. In case of serious content problems, an e-mail can be sent to rrm@relan.at and in case of problems of installation an e-mail to support@relan.at.

Apart from the RELAN-EXE file, all files are text files (.txt) for which the fonts Courier New, Century Gothic or Lucida Console are recommended. At font size point 8, most issues fit in width on an A4 page.

OUTPUT

Subsequently, the most important outputs should be briefly explained, for which it could be helpful to open the file RELAN-OUT.

At the beginning of each output file, you will be informed whether the TRIAL CONDITION (free version) or the LICENSE CONDITION is running.

Between the lines at the beginning of the RELAN-OUT file are the parameters of the input file RELAN-IN, which can be copied, changed, and entered into a new RELAN-IN:

Start of Copy of Input-File: RELAN-IN ***

--- Input parameters ---

End of Copy of Input-File: RELAN-IN ***

CHECK OF VARIABLE NAMES

Check whether correct variable names were used in the formulation of the hypothesis or extraction relation.

CONVERSION OF RELATIONS

Transformation of the three-digit variable names of the hypothesis and the extraction relation into shorter variable symbols (A, B, C, ...).

CHANCE MODELS

Output of which chance models (a priori/aposteriori) were used in the statistical calculations for the hypothesis and for the extraction.

ORDER AND INDICES OF VARIABLES

Output of the selected causal indices (CAUSAL ORDER, CAUSAL DURATION) for the variables. In the illustration CAUSAL RANGE OF VARIABLES, the temporal beginning of a possible causal effect of individual variables as well as their defined effect duration is graphically illustrated.

DATA FILE

Name of the data file used for the evaluations (RELAN-DAT or RELAN-KAT).

DATA VALUES

With the highest value (= 3) of the parameter PRINT EXTENT, all calculated results are output in RELAN-OUT.

INFORMATIONAL ANALYSIS

Information analysis is performed on only ten (integer) variables. The following characteristic values are output: case frequencies in the variable categories (max = 30), information content for cases and variables, total information for the original data and the dichotomized data. The most important result is the ratio of the total information for the original data and that for the dichotomous data, which shows the loss of information caused by the dichotomization of the data.

MINTERM FREQUENCIES

The minterms represent all possible combinations of values of the variables involved. Ordered by the minterms, the frequencies of cases that occur with the corresponding combinations of values are printed.

APOSTERIORI FREQUENCIES

These are the relative frequencies with which the positive values of each variable occur in the sample.

SIMULATION ANALYSIS and **CONFIRMATION ANALYSIS**

CHARACTERISTICS OF THE BOOLEAN FORMULA

Here, the variables contained in the formula are checked and given in abbreviations.

TRANSMISSION OF THE FORMULA

Representation of the successive transformations, in which way the components of the hypothesized relation are calculated to subrelations and finally lead to an overall relation.

TRUTH FUNCTIONS (Variables + Subrelations)

This table shows the logical values of all variables, all resulting subfunctions, and the hypothetical or simulated relation.

PROBABILITIES OF CHANCE MODELS

Probabilities for the chance models (a priori, posterior), which are defined by the relative frequencies of occurrence of positive variable values.

ELEMENTARY RELATIONS

The statistical characteristic values are calculated for all minterms (combinations of variable values) (see RELAN Statistics). The significant minterms are called elementary relations ("types" or "antitypes" in Configuration Frequency Analysis, KFA).

INTERCORRELATIONS

For the estimation of the logical similarities between the variables, subrelations and the hypothesis or simulation relation, product-moment correlations are calculated over all given minterms, on the one hand for the truth values of the relations and on the other hand for their minterm frequencies. Low similarity of a relation with other relations is an important indication of its specific function in explaining the data.

SUBRELATIONS AND HYPOTHETICAL RELATION

This table summarizes the most important statistical parameters for all variables, all subrelations and for the simulated or hypothetical relation (see RELAN Statistics). These statistical parameters for all subrelations derived from the decomposition of the hypothetical relation should take into account by interpretation of the results.

HYPOTHETICAL RELATION

This table contains 30 significance and effect statistics for the simulated or tested relations (see RELAN Statistics).

EXTRACTIONAL RELATION

If the hypothetical or simulated relation is to be reduced by another relation (confounding variables or error relations), then the characteristics

for both the extraction relation and those of the residual relation (reduced relation) are subsequently output.

HIGHER ORDER RELATIONS

The evaluation procedures SIMULATION and CONFIRMATION are concluded by extensive exploratory analyses ("data mining"), in which it is tested for all variables whether "higher order" significant relations are also present between the given variables. These are elementary relations that come about as a function of a smaller set of variables than are given in the tested or simulated relation. In this way, variable relations can be found to which at least one predictive function for a data structure can be attributed (RELAN-OUT-Confirmation-1). (This output is output only at Print Extent = 3.)

SIGNIFICANT HIGHER RELATIONS

Summary output of all significant higher-order relations with associated statistics (see RELAN-OUT-Confirmation-1). (This output is output only at Print Extent = 2 and 3.)

IMPLICATION ANALYSIS and CONNECTION ANALYSIS

These analyses are exploratory and have the purpose of checking predicate logical relationships for significance between two variables. For this type of evaluation, the parameter INCLUSION CRITERION must be set to either "not normalised" or "normalised". In the first setting, the significance values for a relation are calculated in the usual way, in the second they are normalised as if they came about with an equal number of positive and negative truth values of a relation. Thus, relations with different truth values (e.g., conjunctions and implications) can be compared more fairly. By means of the additional parameters Y(p) and/or X(p) it is possible to limit the outputs so that only relations with higher relative frequencies of the two variables are output. In addition, the CAUSAL ANALYSIS parameter can limit the output to only those relations between two variables that match the causal time constraints.

IMPLICATIONS/CONNECTIONS

Only those implications or connections that prove to be significant depending on the parameter setting are output (see RELAN-OUT Implications-2, RELAN-OUT Connections-2). In total, forty statistical parameters are calculated for each significant relation (see RELAN Statistics).

ANALYSIS OF GRAPHS

If this parameter is selected, all significant relations between pairs of variables are subjected to graph analysis (see RELAN-OUT Implications-2). This means that "paths", "paths", "components", "trees" and "spanning trees" are searched in the variable system, on the basis of which starting points, endpoints, branches, connections can be found in a variable network (see RELAN Statistics).

REFERENCE

Maderthaner, R. (2022). RELATIONSANALYSE (RELAN) - Aussagenlogische, statistische und kausale Analyse von Daten. Springer Spektrum, Heidelberg.