Journal for Person-Oriented Research
2015, 1(1-2), 87-98
Published by the *Scandinavian Society for Person-Oriented Research*
Freely available at http://www.person-research.org
DOI: 10.17505/jpor.2015.09

# ROPstat: A General Statistical Package Useful for Conducting Person-Oriented Analyses

**András Vargha[1*], Boglárka Torma[2], Lars R. Bergman[3]**

[1] Institute of Psychology, Károli Gáspár University of the Reformed Church in Hungary, Bécsi út 324, Budapest, Hungary, H-1037
[2] IT Services Hungary Kft, Neumann János u. 1/c, H-1117 Budapest,
[3] Department of Psychology, Stockholm University, 106 91 Stockholm, Sweden

## Email addresses:

vargha.andras@kre.hu, tormabogi@gmail.com, lrb@psychology.su.se

## To cite this article:

Vargha, A., Torma, B., & Bergman, L. R. (2015), ROPstat: A general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research, 1*(1-2). 87-98. DOI: 10.17505/jpor.2015.09

**Abstract:** ROPstat is a wide scope statistical program package which offers specialties in three domains: 1) robust techniques, 2) ordinal analyses, and 3) pattern and person oriented methods. Many of them are not available in other common statistical softwares. In the present paper, first the general features and the main structure of ROPstat are briefly outlined, followed by a more detailed summary of pattern-oriented methods (detecting and imputing missing values, residual case identification, different types of classifications, post-analyses after classifications, etc.). In the last section we present some selected person-oriented scientific questions and show with real-life research data how they can be analyzed using ROPstat.

**Keywords:** ROPstat, statistical software, person-oriented methods, classifications, density variables, core points

ROPstat is a wide scope statistical program package which – beyond the standard toolkit of basic statistical methods – offers specialties in three domains:

- Robust techniques (R)
- Ordinal analyses (O)
- Pattern- and person-oriented methods (P)

Many of them are not available in other common statistical softwares.

The Hungarian ROPstat statistical package was designed by András Vargha who also wrote the computer programs for the different modules. Péter Bánsági contributed to designing the screen interface, and Lars Bergman helped as a consultant with certain person-oriented modules.

In the present paper, first the general features and the main structure of ROPstat are briefly outlined, followed by a more detailed summary of pattern-oriented methods. Finally, we present some selected person-oriented scientific questions and show how they can be analyzed using ROPstat.

## How to Install ROPstat?

ROPstat is a Windows-based program that can be installed from the site of ROPstat, www.ropstat.com. After a successful installation of the demo version, the necessary files of ROPstat will all be copied into a new folder named c:_vargha\ropstat. To launch ROPstat, run ropstat.exe from within this folder. The demo version is already a totally working version of ROPstat that has only two restrictions: it can handle data files with at most 5 variables and 500 cases (larger files will be cut to this size). To obtain the full version the first author should be contacted.

## Data Import and Export in ROPstat

ROPstat creates and reads directly data files of its own format, called msw files. An msw file is a simple text file where both variable characteristics and the data themselves are stored in a special format. In addition, if you choose the appropriate file type in the File/Open menu, ROPstat can

open SPSS files in portable format (*.por) and Excel files in Tab-delimited (*.txt) format. In higher versions of SPSS (21 and above), saving in portable format can only be performed if 'unicode character set' is changed to 'local character set' (in the Edit/Options/Language menu, or in the Edit/Options/Language menu) before opening the SPSS file to be converted to portable format. In the case of Tab-delimited files, the first row must always contain the variable names (preferably not longer than 8 characters). The variables can be of either numeric or text type, but ROPstat can perform statistical analyses only with numerically coded variables.

In order to export data into SPSS or Excel, use the 'Save as in SPSS format' or the 'Save as in Tab-delimited format' choice in the File menu.

## The Menu System of ROPstat

The menu system of ROPstat is based on the natural problem types that occur in practice instead of a statistically minded system. The items of the first three menu choices of the statistical analyses in ROPstat are summarized in Table 1. The fourth menu choice (Pattern-oriented analyses) will be detailed later.

Table 1.
*The items of the first three menu choices of the statistical analyses in ROPstat.*

**1. Basics: single sample analyses**
Basic descriptive statistics
Detailed statistics
Frequency, histogram
Tests on the mean/median
Test of normality

**2. Comparison of groups and variables**
One-way comparison of independent samples
One-way comparison of repeated measures
Two-way independent samples ANOVA
Two-way mixed ANOVA
Two-way independent samples rank ANOVA
Inter-rater reliability for quantitative scales

**3. Relationship of variables**
Correlation, regression
Correlation, partial correlation matrix
Relationship of discrete variables (Crosstab)
Item analysis
Multiple linear regression

The simplicity and user-friendly features of ROPstat can be illustrated by means of a hypothetical data sample presented in Table 2. After starting the demo version of ROPstat, choose 'New File' under 'File'. In the dialogue box that then appears, enter 5 after 'Number of columns (variables)' and 10 after 'Number of rows (cases)'. The data in Table 2 (without variable names) can now be directly copied in the standard way (with the well-known Windows copy/paste operation) to ROPstat. Another option is to copy this table (with variable names) to Excel, save in Tab-delimited text format, and open in ROPstat. In both cases variable characteristics (short and long variable names, definition of group labels, missing values, etc.) can be chosen after shifting from the 'Data Matrix' window to the 'Variables' window in the lower left corner of ROPstat.

Table 2.
*A simple hypothetical data sample with 5 variables and 10 cases.*

| ID | Gender (1: male, 2: female) | Dominance | Anxiety | Femininity |
|---|---|---|---|---|
| 1 | 1 | 16 | 4 | 10 |
| 2 | 1 | 13 | 7 | 13 |
| 3 | 1 | 17 | 6 | 10 |
| 4 | 1 | 10 | 12 | 11 |
| 5 | 1 | 13 | 12 | 8 |
| 6 | 2 | 10 | 18 | 13 |
| 7 | 2 | 11 | 11 | 15 |
| 8 | 2 | 16 | 14 | 14 |
| 9 | 2 | 10 | 7 | 13 |
| 10 | 2 | 10 | 6 | 16 |

With this data sample we can now explore gender differences in terms of the three personality characteristics. To do this, choose 'Statistical Analyses', and then 'Comparing groups or variables', and 'One-way comparison of independent samples'. Then send Gender to the window of 'Grouping variable' and the remaining variables (Dominance, Anxiety, and Femininity) to the window of 'Dependent variables'. Leaving unchanged the default setting of variable scale types (interval) and clicking on 'Run' will perform the comparison of males and females in terms of means and variances for each of the selected dependent variables and the results will appear in an output window. A convenient option in ROPstat is that the content of the output window can be sent to Excel. The results are basically the same as the ones obtained in other well-known statistical programs (descriptive statistics for the groups, check of normality, test of variance homogeneity, effect size

measures, *t*-test and Welch's robust *t*-test). For this reason we do not go into details here.

If the scale type of a dependent variable is changed from *interval* to *ordinal*, the null hypothesis of stochastic equality, equivalent to the equality of expected rank means, will be tested instead of the equality of population means, by means of the Mann-Whitney rank test, and two robust rank tests (see Vargha & Delaney, 2000; Delaney & Vargha, 2002). The results include descriptive rank statistics for the groups, tests of variance homogeneity, a conventional and two robust rank tests, and an ordinal effect size measure, the *A* measure of stochastic superiority (Vargha & Delaney, 1998; 2000). Table 3 shows an excerpt of the summary of the results that are displayed in the output window when the ordinal analysis option is used to analyze gender differences for the Table 2 data set.

Table 3.
*An excerpt of the summary of results from the output of the group comparison module in ROPstat when the scale type of the dependent variables is ordinal for each variable*

| Variable | Rank-mean1 | Rank-mean2 | A_est | p/Levene | p/MW | FPW | p/FPW | BM | p/BM |
|---|---|---|---|---|---|---|---|---|---|
| Dominance | 6.8 | 4.2 | 0.76 | 0.938 | 0.206 | 1.505 | 0.171 | 1,548 | 0.160 |
| Anxiety | 4.1 | 6.9 | 0.22 | 0.456 | 0.183 | -1.664 | 0.145 | -1,710 | 0.136 |
| Femininity | 3.4 | 7.6 | 0.08 | 1.000 | 0.040* | -4.587 | 0.002** | -5.093 | 0.001*** |

*: $p < .05$ **: $p < .01$ ***: $p < .001$

Note. A_est = Sample measure of stochastic superiority; MW = Mann-Whitney *U* test; FPW = Fligner-Policello test with Welch-like *df;* BM = Brunner-Munzel test

A brief summary of the results at the end of the output helps the user of ROPstat to see the most important results. In the present case, Table 3 contains an excerpt of this part of the output. For instance, from Table 3 one can conclude that females are significantly more feminine than males. The $A = A_{12} = 0.08$ value of stochastic superiority indicates a female dominance of Femininity, that is expressed in $A_{21} = 1 - A_{12} = 1 - 0.08 = 0.92$. This value indicates that, if we compare two randomly selected males and females in terms of their Femininity values, we can expect a male dominance (a greater Femininity value) with a chance of about 8%, and a female dominance with a chance of about 92% (in the discrete cases the chance of equality is halved between the two groups). Based on Table 1 of Vargha and Delaney (2000), this level of difference ($A_{21} = 0.92$) corresponds to an effect size greater than Cohen's $d = .80$, which is the lower threshold for large effect size values according to Cohen (see Cohen, 1992, Table 1).

# Pattern-Oriented Analyses in ROPstat

In this section we provide short summaries of the pattern-oriented analyses that can be made in ROPstat. Then we introduce some person-oriented scientific questions and show how they can be analyzed using ROPstat.

With the first three modules (Description, Imputation, and Residue) one can perform preparatory analyses before performing a certain type of classification.

**Description**

This module gives descriptive information of a data set focusing on the configurations of missing values. The output includes basic descriptive statistics, pairwise correlations (optional), and reports about dropouts (for single variables and pairs of variables). An excerpt of a typical output of this module is seen in Table 4.

For each selected variable, Table 4 contains basic descriptive statistics, the number of missing values, and the increase in the number of listwise complete cases if the variable is removed. For example, if Mathgr3 (Math grade 3) would be omitted from the list of selected variables, the number of complete cases would increase by 84 cases. In addition, the table indicates that for 477 cases no variable values are missing (complete cases), for 103 cases exactly 1 variable value is missing, etc.

**Imputation (Imputation of missing values)**

This module performs imputation of missing values by means of three different methods: (1) Replacement with the variable mean; (2) Replacement with the most similar complete case (twin/nearest neighbor); (3) Replacement with the multiple linear regression estimation based on the selected variables. For more details about method (2) see Bergman, Magnusson, & El-Khouri (2003, pp. 107-111).

Table 4.
*An excerpt of the results from the output of Description module of ROPstat.*

BASIC DESCRIPTIVE STATISTICS

| Index | Variable | N_used | Mean | SD | Min | Max | Missings | Missings reduction |
|---|---|---|---|---|---|---|---|---|
| 1 | Gender | 684 | 1.494 | 0.500 | 1 | 2 | 0 | 0 |
| 3 | Mathgr3 | 499 | 3.653 | 0.780 | 2 | 5 | 185 | 84 |
| 5 | Mathgr6 | 588 | 3.753 | 0.827 | 1 | 5 | 96 | 1 |
| 7 | Mathgr8 | 684 | 3.073 | 1.011 | 1 | 5 | 0 | 0 |
| 9 | Mathgr9 | 666 | 3.080 | 1.009 | 1 | 5 | 18 | 9 |
| 16 | Fameduc | 577 | 4.641 | 1.829 | 1 | 7 | 107 | 9 |

*Missing reduction is the number of listwise complete cases added if a variable is removed.*

Number of cases with missing values in exactly K variables

| K variable missings | Number of cases |
|---|---|
| 0 | 477 |
| 1 | 103 |
| 2 | 12 |
| 3 | 89 |
| 4 | 3 |

**Residue (Residue analysis)**

This module identifies residual cases (outliers) based on the distance from the most similar case (cases), called nearest neighbors, and creates a residual indicator variable, by means of which these cases can be omitted from a later classification analysis (using this variable as a conditional grouping variable). To identify the number of outliers, the distribution of all observations of the distance from thenearest (or 2nd or 3rd … nearest) neighbors is created. From this table the number and proportion of cases falling above a specified distance threshold (default = 0.70) can easily be identified (for an example, see Table 6). The applied distance measure is the average squared Euclidian distance (ASED). This means that in the default setting (assuming standardization of the variables) a case can be regarded as outlier if its data values (for the selected variables) differ in absolute value from those of its nearest neighbor on the average by 0.8366 (= square root of 0.70) SD units. The output lists the case indices for all outliers and their distances from their five nearest neighbors.

This module can also create and save up to five density variables with different neighborhood sizes. A density score of a case is related to the number of neighbors the case has, as defined by a certain neighborhood threshold value. To compute the five density scores of a case, first the average number of neighbors for each threshold value is computed. Then, the density score for a case is obtained by dividing the number of neighbors for this case, i.e. the number of observations being closer to the case than the threshold value, by the average size of neighborhood for all cases (for the same threshold value). For example, a density score of 1.65 for a case indicates that the number of neighbors of this case is 65% higher than the average for the given neighborhood threshold value.

The set of density values for a threshold level can be regarded as the altitude value on a geographical map. With an appropriate graphical program (for example ImageJ, see http://imagej.en.softonic.com/download), this can be illustrated in a 3-dimensional picture showing high peaks and large plains in the multidimensional data set. For a discussion of the same type of density variables and their use in cluster analysis, see also Azzalini & Torelli (2007).

One can focus also on the peaks. Setting a cut-off in the scale of a density variable (say cases with density score above 2.0) one can define core cases directly. These can be regarded as dense points (or core points) by means of which dense point regions can be identified in the data set.

With the next four modules (Hierarchical, K-means, DensePoint and CFA) one can perform different types of classification analysis.

**Hierarchical (Hierarchical cluster analysis)**

This module performs an agglomerative hierarchical cluster analysis. Also, it can be followed by an optional relocation analysis (K-means clustering). The user can choose from two distance types of cases (ASED and Pear-

son *r*) and seven different distance types of clusters (Average linkage, Single linkage/Nearest neighbor, Complete linkage/Furthest neighbor, Centroid, Median, Ward, Beta-Flexible). The program computes several summary measures (Explained error sum of squares percentage = EESS%, Point-biserial correlation, Silhouette coefficient, homogeneity coefficients of the obtained clusters, etc.) and creates tables that help to understand the obtained cluster structures. For more details, see the example below of an illustrative real-life example (the results are summarized in Tables 7 to 9), or Bergman, Magnusson, & El-Khouri (2003, Chapter 4, and pp. 113-115).

**K-means (Relocation: nonhierarchical K-means cluster analysis)**

This module improves a cluster solution by the relocation of cases. It starts from an initial classification, and moves cases from one cluster to another if this leads to a reduction in the total error sum of squares of the cluster solution. In this way, bad-fitting cases are moved to a better fitting cluster and more homogeneous clusters can be obtained. The initial classification can be performed by the specification of a starting clustering variable. If such a variable is not given, the program performs automatically a hierarchical clustering with Ward's method stopping at the specified value *k* of the number of clusters that is followed by the relocation process.

In this module the user can also test the significance of a cluster solution by data simulation. Choosing this option the program randomizes the data values in each variable column, performs a K-means clustering with these random data, and computes the explained ESS% corresponding to the obtained cluster solution. Repeating this procedure several times (default = 5) the program tests via Student's *t*-test whether the explained ESS% corresponding to the classification of real data is significantly greater than those obtained from randomized data. For more details about the relocation analysis, see Bergman, Magnusson, & El-Khouri (2003, Chapter 4).

If one is interested in types represented by highly homogeneous clusters, one may perform a cluster analysis with core cases that can be identified by means of density variables (see Residue analysis above). After performing such an analysis, and having obtained an attractive cluster solution of the core cases, one may also be interested in classifying – some or all – non-core cases according to the obtained clusters. This task can be performed by specifying a starting clustering variable. The value of this clustering variable must be set to zero for cases that are intended to be sorted to clusters they are closest to, provided its distance is less than a specified threshold value (so called saving threshold). If there is at least one case for which the value of the starting cluster code variable is zero, instead of performing a K-means cluster analysis, the program will for each such case search for the nearest cluster to it and list it in the output. The distance from a cluster is defined as the distance from the nearest neighbor within this cluster. The optionally saved cluster code variable will include only the cluster code of those cases for which the distance from the nearest neighbor does not exceed the specified saving threshold value. Having a list of centroids of a cluster structure it is also very easy – with this option – to distribute (classify) a set of cases, sorting each case to the nearest centroid.

**DensePoint (Dense point analysis)**

This module searches for cases with many neighbors. A dense point (DP) is defined as a vector (a multidimensional point like a centroid) that has many neighbors. A DP can either be the data list (a vector) of a case for the selected variables, or the center (vector mean) of some cases that are very close to each other. The program first merges iteratively different couples of DP's that are very close to each other to a common DP center (the centroid of such a set), using a certain threshold value (Threshold_1) that can be specified in the panel of the module. Then the neighborhoods of the DP's are identified using another, obviously larger, threshold value (Threshold_2) that can also be specified. If a control clustering variable is specified as well, the program reports the percentages of dense point neighborhoods falling into the different clusters for each explored dense point.

Running the DensePoint module, several useful variables can be created and saved as well. By means of these variables it is easy to identify cases falling into or very close to DP centers (core cases), and their neighbors.

(1) DPcode: For cases closer to a DP less than Threshold_2 (default = 0.25) the variable indicates the index of the nearest DP. As an example, if DPcode = 7 for a case, it means that this case is closest to DP_7.

(2) DPdismin: For cases closer to a DP less than Threshold_2 DPdismin shows the distance of the case to the closest DP (minimal distance from a DP). Cases with very small values of DPdismin (say less than .05) can be regarded as core cases, cases with larger but still small values (say DPdismin < .25) can be regarded as neighbors of core cases. The number of DP's can be controlled in ROPstat by setting the maximal number of DP's (default = 15). For more details, see Vargha & Bergman (in prep.).

Dense points (or core points) can also be obtained in other ROPstat modules as well. Cases with high scores on a density variable created in Residue can also be regarded as dense points. Also if one analyzes the cluster structure of a hierarchical cluster solution with very many – say 100 – clusters one will certainly find some very homogeneous clusters (with homogeneity coefficients less than 0.20) of respectable sizes that can also be regarded as dense points or dense point regions. About dense points, see also Bergman & El-Khouri (2001) or Wishart (1987).

**CFA (Configural frequency analysis)**

In this module configural frequency analysis is performed for identifying value patterns of a small set of discrete variables using the exact binomial test, and applying also a method for probability adjustments (due to Holm) to prevent alpha inflation. Tests are made for both types and antitypes. For more details about CFA see von Eye (2002) or von Eye, Mair, & Mun (2010).

As an illustration of CFA, we describe in Table 5 some results from a study on typical flow patterns during gameplay, at school, and at home, carried out on 1360 11-17 years old Hungarian adolescents (Smohai et al., 2013). The pattern consisted of three variables. Each was coded with 1, 2, 3, 4 or 5, where 5 = very high flow, 4 = high flow, etc. From Table 5 one can see that in the significant configuration patterns (all are types since observed frequencies are greater than expected ones) the values in the pattern are always almost of the same magnitude (111, 222, 333, etc.). This shows that among Hungarian adolescents the flow level is rather stable across different situations. If a teen-ager can feel high or low flow in a game play, he/she can feel the same level of flow also at school or at home in another activity. Though it is very rare that no antitypes occur if several strong types can be identified, in this case this happened. The smallest value of adjusted *p* of antitypes in this analysis was 0.141 (for pattern 253), far from being significant.

Table 5.
*Significant flow patterns during gameplay (1st value under Configuration), at school (2nd value), and at home (3rd value). The table contains an excerpt of the results from the output of the CFA module of ROPstat. For each situation, flow was measured on a 5-point scale (1 = very low flow. 5 = very high flow).*

| Configuration | Observed frequency | Expected frequency | Z-value | Binomial probability | Adjusted *p* | Type (T)/Antitype (A) |
|---|---|---|---|---|---|---|
| 111 | 45 | 5.47 | 16.716 | 0.000 | 0.000 | T |
| 121 | 26 | 6.53 | 7.440 | 0.000 | 0.000 | T |
| 221 | 23 | 8.00 | 5.140 | 0.000 | 0.001 | T |
| 222 | 36 | 9.47 | 8.485 | 0.000 | 0.000 | T |
| 333 | 56 | 13.57 | 11.440 | 0.000 | 0.000 | T |
| 442 | 31 | 12.45 | 5.141 | 0.000 | 0.000 | T |
| 444 | 49 | 9.50 | 12.694 | 0.000 | 0.000 | T |
| 555 | 96 | 18.62 | 17.940 | 0.000 | 0.000 | T |
| Total | 1360 | 1360 | | | | |

With the next four modules (ExaCon, Centroid, Time separation, and Time fusion) one can perform post analyses after classifications.

**ExaCon (Cell-wise analysis of contingency tables)**

In this module exact cell-wise analysis is made of two-way frequency tables. For each cell, it is significance tested whether the cell frequency is higher than expected by chance (a type) or lower than expected by chance (an antitype), based on an exact test (preferred are usually the two-tailed hypergeometric probabilities). With this method the relationship between different cluster code variables, or cluster code variables and other categorical variables (like gender, education level, diagnosis, etc.) can be analyzed in a cell-wise manner. In addition, the degree of similarity between two cluster structures can be assessed by the Rand coefficient (Rand, 1971) or the Adjusted Rand coefficient (Hubert & Arabie, 1985). For more details, see the real-life example in the next section of our paper, or Bergman, Magnusson, & El-Khouri (2003, Chapter 5, and pp. 125-127).

**Centroid**

Centroid can perform a comparison of two classifications. Clustering solutions can be compared with each other by matching each cluster centroid in one solution to the cluster centroid which resembles it the most in the other solution. The outcome is a set of pairs of centroids, each belonging to one solution where the pairs are given in order of decreasing similarity. This module can also compare clustering solutions represented by their centroids (mean patterns) belonging to different subgroups (say males and females). For more details, see the real-life example below, or Bergman, Magnusson, & El-Khouri (2003, Chapter 5, and pp. 125-126).

**Time separation**

This module can be used when a study involves repeated measurements of the same set of variables. A file is created where data at different time points are treated as subindividuals. By means of this method the data record of one case (one row in the data matrix) is divided into separate

parts, one for each measurement occasion, creating several rows in a new file. With this transformation of the original data file interesting classification analyses (such as I-states-as-objects-analysis: ISOA, see Bergman & El-Khouri, 1999; or Bergman, Nurmi, & von Eye, 2012) can be performed.

**Time fusion**

With this module one can reconstruct the original file format (one row for each case) from a time-separated file. A typical task that can be performed by this and the previous module is as follows.

1. Suppose that we have data on the same set of variables (personality scales, IQ-measures, etc.) from the same subjects at different time points.
2. Using module 'Time separation' the variable values belonging to different time points will be placed into different rows in a new data file and hence will be treated as different cases (so called subindividuals). Accordingly, if there are, for example, three different time points, the new, time-separated data file will contain three times as many cases as the original.
3. In the time-separated file, several different analyses can be performed, such as a cluster or factor analysis on the set of the variables in focus, and new variables (cluster code variable, principal components, rotated factors, etc.) can also be created.
4. Finally, we can apply the module 'Time fusion' for recreating the original case file from a time-separated file. Then, this file will contain the values of the newly created variables (e.g., a cluster membership code will be added).

# An illustrative real-life example of person-oriented analyses: Personality patterns among high school teachers in terms of burnout and existential fulfillment

Choosing a job or vocation nowadays does not mean a life-long commitment. Many leave their firstly chosen career path, whereas others remain but without real commitment. However, there are still those who feel real full-term satisfaction with their chosen vocation. In a study of high school teachers we were interested in exploring some psychologically meaningful types related to burnout and existential fulfillment at work.

The participants were 280 Hungarian high school teachers (females: 66.4%, males: 33.6%) from the eastern region of Hungary. Their mean age was 39.6 years (SD = 9.4) and the mean time of working as teacher was 11.2 years (SD = 8.2). For less than half (41.8%) of them the current workplace was the first one.

To achieve the aim formulated above the following variables were chosen.

**Stress**. To assess a global level of stress at work, we used the Effort/Reward ratio from the Effort-Reward Imbalance at Work Questionnaire (Siegrist, 1996).

**Burnout**. To assess a global level of burnout, the sum of the three subscales, Emotional exhaustion, Depersonalization, and Lack of personal accomplishment from the Maslach Burnout Inventory (Maslach & Jackson, 1981) was used.

**Coping**. To assess a global level of coping, we used the Total coping score from Oláh's Psychological Immune System Questionnaire (Oláh, 2005).

**Fulfillment**. To assess a global level of satisfaction with job we used the Existential Fulfillment Scale, a summary score of the subscales Self-acceptance, Self-actualization, and Self-transcendence (Loonstra, Brouwers, Tomic, 2007, 2009).

**Flow**. To assess a global level of Flow, we used the Flow scale from Oláh's Flow Test (Oláh, 2005).

To identify main personality types based on the above listed five variables, a cluster analysis (CA) of cases seems to be a good method. Below we show how this analysis can be performed in ROPstat.

**Preparatory steps**

It is often useful to run the module 'Description' to have some information on the descriptive statistics, the structure of missing values, and the pairwise correlations of the selected variables. Without going into details we note that in the present case we have 204 complete cases for the above defined set of variables, and that the last three (Coping, Fulfillment and Flow) show high pairwise correlations (between .50 and .70), whereas the other correlations are all lower than .15.

If you run CA with the whole sample it may occur that some participants with extreme data (outliers) create a serious bias in the cluster structure. Thus before running CA it is advisable to perform a residue analysis with the specified variables (see Bergman, Magnusson, & El-Khouri, 2003, pp. 109-110), by means of which these outlier cases can be identified and removed. This is accomplished by 'Residue' in the 'Pattern-oriented analyses' menu.

The most important part of the results in 'Residue' is the distribution of the distances from the *k*'th nearest neighbor (the value of *k* can vary between 1 and 5 and can be specified by the value of Number of twins). A part of the obtained results with the variables Stress, Burnout, Coping, Fulfillment and Flow is seen in Table 6.

Cases with the largest distances from their nearest neighbors are found in the bottom of Table 6. There was one case for which the distance from its nearest neighbor was 7.585 and two more cases with a distance value greater than 0.70 (the threshold for a case being considered as an outlier). The Cum% = 98.5 value for the distance value of 0.634 shows that the non-outlier cases represent 98.5% of

all complete cases, whereas the outliers represent 1.5% (= 100-Cum%). If, before running the module, one chooses the option 'Save residual indicator as new variable' a new variable is created (with the name Resid in the last column of the data set), that can be used in other analyses to exclude a group of outliers from the analyses. Here it is worth noting that with this option not only an outlier indicator can be created but five density variables as well. These latter variables show the other side of the story, because they make it possible to identify cases with many close neighbors, the most central cases in the data set with respect to the selected variables. The five density variables correspond to five increasing sizes (radius values) of neighborhood around the cases.

Table 6.
*An excerpt of the results from the output of the Residue module of ROPstat (from the table of distribution of distances of nearest neighbors)*

| Distance | Frequency | % | Cum % | 100-Cum % |
|---|---|---|---|---|
| 0 | 3 | 1.5 | 1.5 | 98.5 |
| 0.010 | 4 | 2 | 3.4 | 96.6 |
| 0.014 | 2 | 1 | 4.4 | 95.6 |
| … | … | … | … | … |
| 0.600 | 1 | 0.5 | 98 | 2 |
| 0.634 | 1 | 0.5 | 98.5 | 1.5 |
| 0.723 | 1 | 0.5 | 99 | 1 |
| 1.896 | 1 | 0.5 | 99.5 | 0.5 |
| 7.585 | 1 | 0.5 | 100 | 0 |
| Total | 204 | | | |

**Classification analyses**

After the preparatory steps we can perform classification analyses. We start with a hierarchical CA (the module 'Hierarchical'). Because our five variables have interval scales, the default setting (Euclidian measure for the distance of cases with Ward type fusion of clusters) seems to be appropriate. In the first run we are interested in the list of the most important adequacy measures (EESS%, Point-biserial correlation, Silhouette coefficient, average cluster homogeneity, etc.). For each step of the hierarchical analysis,

$$EESS\% = 100*(SStotal - SScluster)/SStotal,$$

where SStotal is the sum of the sum of squared deviations from the input variable means for the whole sample (if standardization is chosen, this is performed for the standardized data), whereas SScluster is the sum of the sum of squared deviations from the input variable centroids for each cluster, and summed over the clusters of the actual step. This is a kind of explained variance measure.

The Point-biserial correlation is the usual $r_{XY}$ Pearson-correlation computed the following way. All complete cases are paired with each other. Variable $X$ is a binary variable with a value 0 if the pair of cases belongs to the same cluster and 1 if not. Variable $Y$ is the distance of the two paired cases (ASED). The Point-biserial correlation is high if pairs being in the same clusters are substantially closer to each other than pair of cases that belong to different clusters. The homogeneity coefficient of a cluster is the average of the pairwise within-cluster distances of cases. The Silhouette coefficient is an adequacy measure that takes into account both the homogeneity of the clusters and the level of separation of the different clusters (Rousseeuw, 1987).

In order to choose a suitable number of clusters we first set 1 for both the lower and upper value of the number of clusters for which detailed results will be listed in the output. After running 'Hierarchical' (using standardized variables, and Resid, the residual indicator variable, as conditional grouping variable in order to omit outliers from the analysis), we have the most important results in Table 7.

In Table 7 we look for a $k$ clusters solution where the EESS% is well above 50% and at the next fusion drops substantially to a lower level. In our case $k$ = 5 (with EESS% = 57.30) seems most close to fulfill this expectation and we have a hope that after performing a relocation based on the $k$ = 5 solution we will reach a still higher EESS% level. We then repeat the hierarchical cluster analysis with the following modifications: (1) we set 5 for both the lower and upper value of the number of clusters for which detailed results will be listed in the output; 2) we asked the program to perform relocation after the hierarchical analysis. The obtained results are summarized in Table 8 and Table 9, and can be explained as follows.

The EESS% level increases from 57.30 to 62.43, due to relocation (see Table 8). The main results concerning this cluster structure are reflected by the standardized means (centroids) for each cluster (see Table 9). In the last column of Table 9, tentative short names are provided to characterize the main features of the obtained clusters. The clusters seem to be meaningful, though some are a bit heterogeneous. A rule of thumb is that reasonably homogeneous clusters should have homogeneity coefficients well below 1. Since 2 out of 5 clusters do not fulfill this expectation and the average of the homogeneity coefficients of the five clusters is also fairly large (0.703), we took the six-cluster solution into consideration as well.

Table 7.
*An excerpt of the results from the output of Hierarchical module of ROPstat (from the table of the results of stepwise fusion of clusters) (N = 201)*

| Step | CL# | EESS% | Pointbis | SilCoef | MeanHC | HC range | Clusters to fuse | |
|---|---|---|---|---|---|---|---|---|
| i=0 | 201 | 100 | | | | | 213 (1) | 220 (1) |
| i=1 | 200 | 100 | | | | | 213 (2) | 224 (1) |
| i=2 | 199 | 100 | | | | | 150 (1) | 236 (1) |
| … | … | … | | | | | … | … |
| i=191 | 10 | 70.56 | 0.278 | 0.452 | 0.567 | 0.36-1.12 | 8 (9) | 11 (15) |
| i=192 | 9 | 68.63 | 0.279 | 0.444 | 0.600 | 0.36-1.12 | 1 (25) | 4 (25) |
| i=193 | 8 | 66.53 | 0.311 | 0.457 | 0.636 | 0.36-1.12 | 15 (34) | 29 (18) |
| i=194 | 7 | 63.83 | 0.340 | 0.457 | 0.685 | 0.56-1.12 | 31 (13) | 39 (22) |
| i=195 | 6 | 60.78 | 0.341 | 0.407 | 0.738 | 0.56-1.12 | 15 (52) | 16 (23) |
| i=196 | 5 | 57.30 | 0.382 | 0.436 | 0.800 | 0.61-1.12 | 31 (35) | 72 (17) |
| i=197 | 4 | 53.47 | 0.377 | 0.480 | 0.865 | 0.61-1.27 | 1 (50) | 8 (24) |
| i=198 | 3 | 46.19 | 0.380 | 0.486 | 0.996 | 0.75-1.27 | 1 (74) | 31 (52) |
| i=199 | 2 | 33.91 | 0.349 | 0.616 | 1.215 | 0.75-1.49 | 1 (126) | 15 (75) |
| i=200 | 1 | 0 | 0 | 1 | 1.490 | - | - | - |

Note. CL# = number of clusters; EESS% = Explained Error Sum of Squares %; Pointbis = Point-biserial correlation; SilCoef = Silhouette coefficient; HC = homogeneity coefficient

Table 8.
*An excerpt of the relocation results for five clusters from the output of Hierarchical module of ROPstat (from the table of iterations)*

| Iteration number | Relocated cases | Current ESS | Explained ESS% | Point-biserial correlation |
|---|---|---|---|---|
| i=1 | 42 | 353.48 | 61.39 | 0.384 |
| i=2 | 20 | 348.99 | 61.88 | 0.377 |
| i=3 | 10 | 346.49 | 62.16 | 0.374 |
| i=4 | 9 | 344.51 | 62.37 | 0.376 |
| i=5 | 4 | 343.97 | 62.43 | 0.379 |
| i=6 | 1 | 343.96 | 62.43 | 0.378 |
| i=7 | 0 | 343.96 | 62.43 | 0.378 |

Table 9.
*The pattern of standardized means for the five clusters solution in the original sample (N = 201) after relocation (H = High. L = Low; more extreme levels in either direction are indexed with more pluses; parentheses indicate slighter deviations from the average. and a dot indicates an about average level)*

| Cluster index | Cluster size | Homogeneity coefficient | Stress | Burnout | Coping | Fulfillment | Flow | Short name of cluster |
|---|---|---|---|---|---|---|---|---|
| 1 | 60 | 0.52 | (L) | . | . | . | . | Average |
| 2 | 37 | 0.94 | H++ | . | . | . | . | High stress |
| 3 | 41 | 0.57 | . | L+ | H+ | H+ | H+ | High fulfillment |
| 4 | 41 | 0.72 | . | H | (L) | (L) | (L) | Tendency for burnout |
| 5 | 22 | 1.02 | . | H++ | L++ | L++ | L++ | Severe burnout |

With the six-cluster solution the EESS% level increased to 64.44 after relocation, the average of the homogeneity coefficients decreased to 0.670 (min = 0.48; max = 1.16), and the homogeneity coefficients were less than 1 for 5 out of the 6 clusters. Since this solution still does not seem to be optimal, we decided to omit from the cluster analysis the

peripheral cases. For this reason we again used residue analysis but with a more stringent threshold value (0.20 instead of the default 0.70). For this reduction the number of complete cases decreased only by 15% (from 201 to 170).

Having performed cluster analyses with this reduced sample with 5 and 6 clusters and with and without relocation we summarized the main adequacy measures of the results in Table 10 together with the corresponding results obtained with the original complete sample.

Comparing the adequacy measures of the different solutions in Table 10 one can conclude that KClus5r (K-means analysis on the reduced sample with $k = 5$) seems to be a good solution in several respects: EESS = 66.09 is among the highest values, the Silhouette Coefficient is the highest, and Point-biserial, Mean HC and HC range are near the best values. The pattern of the standardized means for this solution can be seen in Table 11.

Table 10.
*The main adequacy measures of different cluster analyses performed with the original (N = 201) and the reduced (N = 170) sample where peripheral cases were dropped from the analyses*

| CA label | Type of CA | N | No. of clusters | EESS% | Pointbis | SilCoef | Mean HC | HC range |
|---|---|---|---|---|---|---|---|---|
| KClus5 | Hier + Reloc | 201 | 5 | 62.43 | 0.378 | 0.544 | 0.703 | 0.52-1.02 |
| Kclus6 | Hier + Reloc | 201 | 6 | 64.44 | 0.366 | 0.522 | 0.670 | 0.48-1.16 |
| HClus5r | Hierarchical | 170 | 5 | 63.47 | 0.369 | 0.513 | 0.581 | 0.39-0.77 |
| KClus5r | Hier + Reloc | 170 | 5 | 66.09 | 0.363 | 0.569 | 0.539 | 0.36-0.75 |
| Hclus6r | Hierarchical | 170 | 6 | 66.59 | 0.336 | 0.494 | 0.534 | 0.39-0.77 |
| Kclus6r | Hier + Reloc | 170 | 6 | 68.49 | 0.339 | 0.556 | 0.504 | 0.35-0.75 |

Note. CA = cluster analysis; EESS% = Explained Error Sum of Squares %; Pointbis = Point-biserial correlation; SilCoef = Silhouette coefficient; HC = homogeneity coefficient

Table 11.
*The pattern of standardized means for the five clusters solution in the reduced sample (N = 170) after relocation (H = High. L = Low; more extreme levels in either direction are indexed with more pluses; parentheses indicate slighter deviations from the average, and a dot indicates an about average level)*

| Cluster index | Cluster size | Homogeneity coefficient | Stress | Burnout | Coping | Fulfillment | Flow | Short name of cluster |
|---|---|---|---|---|---|---|---|---|
| 1 | 37 | 0.45 | . | . | (L) | (L) | . | Average- |
| 2 | 30 | 0.71 | H++ | . | . | . | . | High stress |
| 3 | 40 | 0.36 | (L) | (L) | . | . | . | Average+ |
| 4 | 30 | 0.49 | . | L+ | H++ | H+ | H+ | High fulfillment |
| 5 | 33 | 0.75 | . | H++ | L+ | L+ | L+ | Burnout |

The centroids of three clusters (with indices 2, 4, 5) in the new cluster solution described in Table 11 are very similar to three clusters from the original five-cluster solution shown in Table 9 (in these cases we assigned the same names to the clusters), whereas for Cluster 1 and Cluster 3 the correspondences are not so clear. Cluster 1 shows some tendency toward negative states (slightly low in coping and fulfillment), so we labeled it Average-, by contrast to Cluster 3 that shows a tendency toward positive characteristics (slightly low in stress and burnout, and thus labeled Average+). To make a more thorough, analytical comparison between the two cluster structures post-analyses are needed that will be presented in the next section.

**Post-analyses**

To further compare the five-cluster structures of the original (KClus5) and the reduced (KClus5r) samples, the Exacon module of the ROPstat was run. From this analysis we obtained global similarity measures reflecting a high similarity (Rand index = 0.844) but also with a substantial difference (Adjusted Rand index = 0.543) between the two cluster structures. This was confirmed also by the bivariate frequency table of the two cluster code variables created by this same module (see Table 12).

Table 12. *The two-dimensional frequency table of the clustering variables of the original (KClus5) and the reduced (KClus5r) samples*

| | Values of cluster code variable Kclus5r | | | | | |
|---|---|---|---|---|---|---|
| Kclus5 | 1 | 2 | 3 | 4 | 5 | Total |
| 1 | 20 | 3 | 33 | 0 | 0 | 56 |
| 2 | 0 | 27 | 0 | 0 | 0 | 27 |
| 3 | 0 | 0 | 7 | 30 | 0 | 37 |
| 4 | 17 | 0 | 0 | 0 | 17 | 34 |
| 5 | 0 | 0 | 0 | 0 | 16 | 16 |
| Total | 37 | 30 | 40 | 30 | 33 | 170 |

Running the module Centroid we obtained a pairwise matching of the centroids of the two cluster structures (see Table 13). In Tables 12 and 13 we see two almost identical clusters ('High stress' and 'High fulfillment') and two rather similar clusters ('Average+' and 'Burnout').

Table 13.
*Pairwise centroid matchings in increasing order of cluster structures KClus5 and KClus5r*

| | Cluster matching | | Distance |
|---|---|---|---|
| Index | Kclus5 | Kclus5r | ASED |
| 1 | 2 | 2 | 0.003 |
| 2 | 3 | 4 | 0.006 |
| 3 | 1 | 3 | 0.058 |
| 4 | 5 | 5 | 0.088 |
| 5 | 4 | 1 | 0.148 |
| 6 | 4 | 5 | 0.189 |

We compared also the cluster structures of Kclus5r and Kclus6r by means of Exacon and Centroid. In terms of similarity measures we obtained an outstanding similarity (Rand index 0.948, Adjusted Rand index = 0.832), that was also confirmed by the pairwise centroid matchings of the two cluster structures. The obtained correspondences were strong for four clusters of the reduced sample (Clusters 1, 3, 4, 5 in Table 11). Cluster 2 ('High stress') was divided into two sub-clusters in the six-cluster solution that could not be interpreted so clearly as the 'High stress' group in the five-cluster solution.

Summarizing the results of these person-oriented analyses, we can draw the following conclusions:

1. The results identified two main personality types among high school teachers with strong but opposite patterns in relation to mental health and their profession: 'High fulfillment' and 'Burnout', which, respectively, represent a successful and an unsuccessful career of teachers.

2. A third well identifiable type includes those teachers whose main symptom is a highly elevated level of stress at work. This group can be regarded as a risk group for which appropriate stress reducing interventions are called for.

3. The last two types (Average- and Average+; see Table 11) might consist of those teachers who have a tendency either toward a positive (fulfillment) or a negative (burnout) condition. These persons may also benefit from support, either for the sake of strengthening the good directions in their personality development or for protecting them from unwanted bad outcomes.

# Discussion

A primary goal of the present paper was to draw the attention of the reader to the pattern-oriented modules of ROPstat that are useful for a wide range of person-oriented analyses. Hopefully, we could convince the reader that a standard pattern-oriented analysis is generally more complex than just running a type of CA.

It is very important before CA to study the pattern of missing data, occasionally to make data imputations, and identify outliers that may seriously distort the classification results.

It is also of primary importance that classifications should be performed with the least level of subjectivity. For this reason one has to evaluate the resulting classifications carefully in terms of the main adequacy measures (EESS%, point-biserial correlation, cluster homogeneity coefficients, etc.) and the interpretability of the obtained cluster centroids. It is helpful to compare some promising cluster solutions in post-analyses via the modules Exacon and Centroid to assess their overall similarities, and finding the pairwise correspondences of the centroids of different cluster structures.

The classification results of the presented study of high school teachers resulted in two different close to average clusters in the final structure (Average- and Average+; see Table 11). These types might be of a transitory nature, ready to develop toward either of two more stable states, fulfillment and burnout. To shed further light on this issue, longitudinal data are needed. Then more complex classification analyses can be performed, enabling the user to test structural stability (similar clusters at different occasions) and individual stability (individuals belonging to similar clusters at different occasions). Two of these complex classification analysis types are LICUR (Bergman, Magnusson, & El-Khouri, 2003) and ISOA (Bergman & El-Khouri, 1999).

# References


Azzalini, A., & Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing, 17 (1),* 71–80. doi: 10.1007/s11222-006-9010-y

Bergman, L.R., & El-Khouri, B.M. (1999). Studying individual patterns of development using I-States as Objects Analysis (ISOA). *Biometrical Journal, 41 (6),* 753-770.

Bergman, L.R., & El-Khouri, B.M. (2001). Developmental processes and the modern typological perspective. *European Psychologist, 6 (3),* 177-186. doi: 10.1027//1016-9040.6.3.177

Bergman, L. R., Magnusson, D., & El-Khouri, B. M. (2003). *Studying individual development in an interindividual context. A Person-oriented approach.* Mahwa, New Jersey, London: Lawrence-Erlbaum Associates.

Bergman L. R., Nurmi J.-E., & von Eye, A. A. (2012). I-states-as-objects-analysis (ISOA): Extensions of an approach to studying short-term developmental processes by analyzing typical patterns. *International Journal of Behavioral Development, 36 (3),* 237-246. doi: 10.1177/0165025412440947

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112 (1),* 155-159.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods, 7 (4),* 485-503. doi: 10.1037/1082-989X.7.4.485

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Loonstra, B., Brouwers, A., & Tomic, W. (2007). Conceptualization, construction and validation of the Existential Fulfilment Scale. *European Psychotherapy, 7 (1),* 5-18.

Loonstra, B., Brouwers, A., & Tomic, W. (2009). Feelings of existential fulfilment and burnout among secondary school teachers. *Teaching and Teacher Education, 25 (5),* 752-757. doi:10.1016/j.tate.2009.01.002

Maslach, C., & Jackson, S. E. (1981). *Maslach Burnout Inventory Manual*. Palo Alto, California: Consulting Psychologists Press.

Oláh, A. (2005). *Anxiety, Coping, and Flow: Empirical studies in interactional perspective.* Budapest: Trefort.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66 (336),* 846–850.

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics 20,* 53–65.

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology, 1 (1),* 27-41.

Smohai, M., Mirnics, Z., Vargha, A. Torma, B., & Tóth, D., (2013). Videojátékokkal való játszás közben, iskolában és otthon átélt flow-élmények tipikus mintázatai, valamint az azokba tartozó magyar serdülők személyiségjellemzői és megküzdési módjai - Konfigurációelemzés. [Typical flow patterns during video game play, at school and at home, and personality and coping characteristics of Hungarian adolescents: a configuration analysis study.] *Pszichológia, 33 (4),* 313-327.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics. 23 (3),* 170-192.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25 (2),* 101-132.

von Eye, A. (2002). *Configural Frequency Analysis - Methods, Models, and Applications*. Mahwah, NJ: LawrenceErlbaum.

von Eye, A., Mair, P., & Mun, E.-Y. (2010). *Advances in Configural Frequency Analysis.* New York: Guilford Press.

Wishart, D. (1987). *Clustan User Manual.* 16 Kingsburgh Road, Edinburgh: Clustan Ltd.