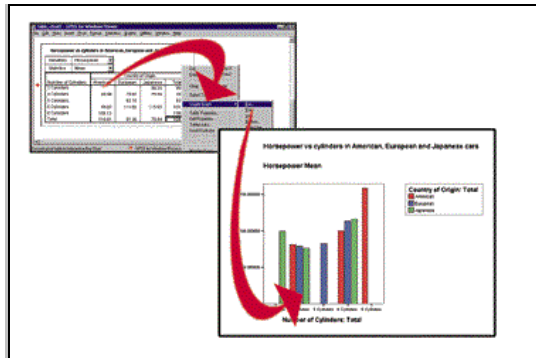


Analysing Environment Indicators with SPSS



Manual supporting the course

compiled by Klaus Röder - may 2002

Table of Contents

- 1. INTRODUCTION 3**
- 1.1. CONVENTIONS OF TEXT STYLES 4
- 1.2. HOW TO CHOOSE OPTIONS FROM MENUS. 4
- 1.3. EXCEL AND SPSS: DIFFERENCES OF USE FOR STATISTICAL PURPOSE 5
- 1.4. USE OF STATISTICAL ANALYSIS IN EXCEL: REGRESSION ANALYSIS..... 5
- 2. A GENERAL VIEW..... 8**
- 2.1. SPSS BASICS.2. SPSS WINDOWS 8
- DATA ENTRY..... 10
- REGRESSION COMMAND IN SPSS. 10
- 4. THE DATA FILE 10
- 2.2. SAVE THE FILES..... 12
- 3. MULTIVARIATE AND COMPLEX STATISTICS 13**
- 3.1. CLUSTER ANALYSIS 13
- 3.2. CALL THE CLUSTER COMMAND..... 15
- FACTOR ANALYSIS 18
- CALL THE FACTOR COMMAND 19

1. Introduction

The aim of this course is to introduce the participants briefly to the statistics program SPSS. This program is among the best known and most widely distributed statistics programs. For many years, since the days of the mainframe, where this program was originally developed, it stands for a very efficient symbiosis between computers and statistical analysis.

Since this course deals with environmental statistics, the special emphasis is on the use of SPSS for environmental statistics.

Since the participants of this course should be familiar with the use of EXCEL, a point of distinction is made between the use of EXCEL for statistical analysis and SPSS. The purpose is NOT to introduce SPSS as a system, but to understand the basic principals of handling this software package and to understand the benefits and advantages against more general software like EXCEL.

Since the course focuses on some statistical methods for analysis, it might be the side effect of this course for some participants to refresh old theory and gain new insight into some elementary statistical methods. The course addresses the beginner to SPSS as well as the experienced SPSS user, it assumes that the participant is equipped with a basic statistical background. The result should be that the participants are familiar with some basic concepts of SPSS and understands the use of some more sophisticated methods to analyze data for environmental statistics.


The development of computers has been dramatic over the recent years, so the software for analyzing and communication of statistical results has been developed as well. Although the statistical theory behind the software has not changed, if the focus remains on the less exotic and sophisticated methodology, the impact on user friendliness and "look and feel" of the software was tremendous.

The actual version 10 of SPSS is fully integrated into WINDOWS as well as network computing and eases the use and access notably for beginners. The Common User Access(CUA) of WINDOWS and the integrated help features facilitates the learning process and helps even experienced users handling the program. The data and information exchange between SPSS and other WINDOWS applications eases and speeds up the work of the analyst.

This introduction will not replace a SPSS reference manual, but the use this manual will accompany the participants through the set of guided exercises and together with the on-line tutorial and the on-line help should be sufficient for the aims of the course

1.1. Conventions of text styles

Throughout this manual we will follow certain conventions. We hope, that this will make the text easier to understand.

[Alt]	You are instructed to type the indicated key
[Alt] + [Shift]	You are instructed to type the indicated keys together
[End],[_]	You are instructed to type the indicated keys one after the other
 [Files/Open/Data]	Menu choices, you can use mouse or keyboard to activate the menu choice

Data output in SPSS presented during the course



Presenting the Data file first in EXCEL then imported to in SPSS

1.2. How to choose options from menus.

There are two methods of choosing items from a menu:

With a mouse: Click on the name of the menu that you want. The name of the menu is highlighted and the list of menu items drops down. You can then view the items and click on the particular item that you want. (Or click on the menu name, holding your finger on the mouse button. Without releasing the button, drag through the menu items in order. They become highlighted in succession as you drag through the list. When the item that you want is highlighted, release the mouse button.) To cancel: click outside the menu and menu bar.

With the keyboard: Press the [Alt] key and the underlined letter (often the first) in the menu name. When the menu drops down, press the underlined letter in the command name. To cancel: press [Esc].



1.3. EXCEL and SPSS: Differences of use for Statistical Purpose

	EXCEL	SPSS
Range of Statistical Tools	Sufficient for a large range of use Sufficient for non-specific purposes	Appropriate for detailed and sophisticated Stats
Data Quantity /Speed	Limited, with Open Database Connectivity (ODBC) theoretically unlimited Speed limited (MS-Windows specific)	Unlimited, but use of Open Database Connectivity (ODBC) recommended for large data sets. Speed limited (MS-Windows specific) better than EXCEL
Tool performance/ Ease of use	Excellent for non-specific purposes and interactive use. EXCEL is a hybrid software for all kinds of purposes, among this statistical analysis	Difficult to use for beginners. This is a tool for statisticians and professionals

1.4. Use of statistical analysis in EXCEL: Regression Analysis

Microsoft Excel provides a set of data analysis tools, called the *Analysis ToolPak* that you can use to save steps when you develop complex statistical or engineering analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

To view a list of available analysis tools, click Data Analysis on the Tools menu.

Regression analysis tool

This analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables ³/₄ for example, how an athlete's performance is affected by such factors as age, height, and weight.

PRESENTATION:

National data set of Series related to environmental topics

Problem:

Are there any indications that here is a relation between infant mortality and environmental series?

EXERCISE:

Use the National Data Set to relate :

Infant Mortality (the independent variable) (SP.DYN.LE00.IN)

and

Fertilizer consumption (AG.CON.FERT.ZS)

and in the second step

CO² emissions (EN.ATM.CO2E.PC)

BACKGROUND:

What is (Linear) Regression:

In many statistical studies, the goal is to establish a relationship, expressed via an equation, for predicting typical values of one variable given the value of another variable. The simplest equation is that of a straight line:

The equation for the line is:

$$y = mx + b \text{ or } y = m^1x^1 + m^2x^2 + \dots + b \text{ (if there are multiple ranges of } x\text{-values)}$$

where

x^n are the independent variables

y is the dependent variable

b is the constant or intercept

m^n are three regression coefficients

Scatterplots are useful for assessing how appropriate a straight line is to summarize the relationship

From scanning the line in a plot, you can see that when literacy is high, so is life expectancy; and when literacy is low, life expectancy tends to be short. The relationship is *linear* because the points scatter fairly evenly around the line.

A linear relation can also occur when the line tilts from the upper left corner down to the lower right. For example, life expectancy tends to go down as infant mortality rate goes up. If the line is flat, extending horizontally across the plot, there is no linear relation. If you see that a straight line summarizes the relationship poorly.

There are several statistical indicators illustrating the 'quality' of the relation between the variables

In EXCEL this is the summary output table, which includes an

- anova table,

- coefficients,
- standard error of y estimate,
- r2 values,
- number of observations
- standard error of coefficients.

The main indicator for the quality , how well one (or many) variable(s) explains the dependent variable are the R and R squared values



Close to zero means : little or no relation
Close to +1 means: strong relation

EXAMPLE:

This Exercise in EXCEL

2. A general view.

- ✓ The objective of this chapter is to give you a general view of the program. You should learn which are the main elements of SPSS, the types of windows and files SPSS uses

How to start and stop SPSS.1. How to start and stop SPSS

2.1. SPSS basics.2. SPSS windows

How to start and stop SPSS2.1. How to start and stop SPSS

- Click two times on the SPSS icon to start
- Choose [File/Exit] to leave the program

SPSS uses the following most important types of windows

Data editor: A rectangular, spreadsheet-like display of the working data file. You can edit the data in the data window, add or delete variables, or change their attributes, except when the SPSS processor is modifying the data. Use File menu commands New and Open to clear data from the data editor, or to open an existing data file in the data editor. Use File menu commands Save or Save As when the data editor is active to save the data file. You cannot close the data editor, although you can minimize it. The file extension used by SPSS for this type of window is .sav.

Viewer. All statistical results, tables, and charts are displayed in the Viewer. You can edit the output and save it for later use. A Viewer window opens automatically the first time you run a procedure that generates output.

Results are displayed in the Viewer. You can use the Viewer to:

Browse results.

- Show or hide selected tables and charts.
- Change the display order of results by moving selected items.
- Move items between the Viewer and other applications.

The Viewer is divided into two panes:

The left pane of the Viewer contains an outline view of the contents.

The right pane contains statistical tables, charts, and text output.

You can use the scroll bars to browse the results, or you can click an item in the outline to go directly to the corresponding table or chart.

You can click and drag the right border of the outline pane to change the width of the outline pane.

Through the Viewer you can call various other editors

- **Pivot Table Editor.** Output displayed in pivot tables can be modified in many ways with the Pivot Table Editor. You can edit text, swap data in rows and columns, add color, create multidimensional tables, and selectively hide and show results.
- **Chart Editor.** You can modify high-resolution charts and plots in chart windows. You can change the colors, select different type fonts or sizes, switch the horizontal and vertical axes, rotate 3-D scatterplots, and even change the chart type.
- **Text Output Editor.** Text output not displayed in pivot tables can be modified with the Text Output Editor. You can edit the output and change font characteristics (type, style, color, size).

Syntax window with Syntax Editor: You can paste your dialog box choices into a syntax window, where your selections appear in the form of command syntax. You can then edit the command syntax to utilize special features of SPSS not available through dialog boxes. You can save these commands in a file for use in subsequent SPSS sessions.

If you have more than one open Viewer window, output is routed to the designated Viewer window. If you have more than one open Syntax Editor window, command syntax is pasted into the designated Syntax Editor window. The **designated** windows are indicated by an exclamation point (!) in the status bar. You can change the designated windows at any time. The designated window should not be confused with the **active** window, which is the currently selected window. If you have overlapping windows, the active window appears in the foreground. If you open a new Syntax Editor or Viewer window, that window automatically becomes the active window and the designated window.

Data Entry

Types of data entry

- Data Entry (Additional package for SPSS)
- Input from EXCEL (LOTUS 1-2-3, ACCESS etc.)
- and use of Forms in example given
EXCEL or
ACCESS or other

Regression command in SPSS.

4. The data file

Open the national data file:

[File/Open/Data]

Then a window is opened with the following name: *BLZ???.SAV*

for the country (e.g. Belize) and ??? for the actual number

Call the regression command

[Analyze/Regression/Linear]

Choose the variables, statistics etc.

Select one numeric Dependent variable, and one or more numeric Independent variables.

You can control the entry of Independent variables into the analysis in two ways: by grouping them into blocks, and by choosing the method by which the variables in each block are processed. The available methods are Enter, Remove, Stepwise, Backward, and Forward. Starting with the first block, SPSS applies the selected method to all of the variables in the block, then proceeds to the next block if there is one.

Select **Statistics** for additional statistics.

Select **Plots** for residual scatterplots, histograms, outlier plots, or normal probability plots.

Select **Save** to create new variables containing predicted values, residuals, and related statistics.

Select **Options** to change the criteria used in the stepwise methods, to request regression through the origin, or to control the treatment of missing value

- Estimates are the coefficients themselves.

- Confidence intervals are 95% confidence intervals for the coefficients
- Covariance matrix gives the variances and covariances among the coefficient estimates.
- Descriptives provides the means and standard deviations of each variable in the analysis, plus a correlation matrix (with one-tailed significance level and number of cases for each correlation).
- Model fit statistics include multiple R, R squared and adjusted R squared, standard error of the estimate, and an analysis-of-variance table.
- additionally
- Durbin Watson displays the Durbin-Watson test for serial correlation of the residuals.

Execute the commands or [Paste] to the syntax editor

Look at the results (View and Output Files)

BACKGROUND

The **slope** is the ratio between the vertical change and the horizontal change at the line. The **intercept** or **constant** as it is often called, is where the line intercepts the vertical axis at $x = 0$ (that is, when life expectancy 0, the intercept is the height from 0 on the expectancy scale to the line).

To represent the errors (e) in the model, draw a short vertical line from each point to the line. The lengths of these line segments between the line and the plot points called residuals and are estimates for the true errors. SPSS uses the method of *least squares* to estimate the slope and intercept. This method minimizes the sum of squared residuals (that is, the sum of the squares of the vertical line segments).

In the first equation, y is the **dependent or outcome** variable, the one you are trying to predict; x is the independent or **predictor** variable; and the intercept and slope are coefficients. If the model is a good descriptor of the relation between the variables, you can use the estimates of the coefficients to *predict* the value of the dependent variable for new cases.

Models with **two or more** predictors. Adding a second independent variable: The equation with one independent variable is the model for **simple linear regression**; the equation with two variables is a model for multiple regression. SPSS allows you to include more than two independent variables in a multiple regression

Correlations. If the Pearson correlation between the two variables is significant with a value less than 0,1 indicating the hypothesis that the correlation is 0 (no linear relation between the variables) is rejected. When your model includes more than one independent variable, SPSS displays correlations for all pairs of variables. In the last panel, labeled N , sample sizes are reported for each variable individually and for each pair of variables.

Normality is not required for the estimates of the coefficients. To make tests and estimate confidence intervals, however, these assumptions are required:

- The errors are normally distributed with mean 0.
- The errors have constant variance.

- The errors are independent of each other.

These assumptions are checked by studying the residuals from the model. The Durbin-Watson statistic (available on the Linear Regression: Statistics subdialog box) can be used to test for the serial correlation of adjacent error terms.

Model Summary. The value of R (also called multiple R) . When there is only one independent variable, R is the simple correlation between the independent and dependent variable (see the Same correlation in the Correlations table above). R^2 is the square of this value and often is interpreted as the proportion of the total variation in life expectancy accounted for by the independent variable (fertilizer consumption “explains“ ???% of the variability of life expectancy). ranges from 0 to 1. If there is no *linear* relation between the dependent and independent variable, R^2 is 0 or very small. If all the observations fall on the regression line, R^2 is 1. This measure of the goodness of fit of a linear model is also **called the coefficient of determination**.

A caution. Be careful about concluding, “If fertilizer consumption is decreased, the population will live longer.“ There is a *association* between fertilizer consumption and life expectancy. However, these data come from an observational study, not a controlled experiment; so any statements about cause-and-effect relationships can be misleading. Association is not the same as causation..

If an investigator observes the values of the independent and dependent variables for a set of subjects (cases), association does not establish causation. If an investigator does an experiment where he or she sets the values of the independent variable (for example, six specific doses of a drug) and watches the effect on the dependent variable, there may be little question that the results were *caused* by the independent variable.

2.2. Save the files

[File/Save As]

If this is a data files, then type the file name, e.g.: *BRZ01* ,the file will be saved as *BRZ01.sav*. Watch where you save your work. It is a good idea to create a separate directory for your data files. You should do this using the WINDOWS-Explorer.

Generally SPSS will save your work in the actual directory, e.g. the SPSS directory where all the system files are stored. This is no good place for your data files or later for your results.

High-resolution graphics allows interpretation of results with more ease. Especially statistical publication and on the spot information for a less professional public asks for extensive use of graphical representation of statistical data.

Save the output file in the SPSS Viewer

1. in the SPSS format (.spo file) *or*
1. export in Internet file format (.htm). The text files will be in .htm format, the graph in JPEG (.jpg) format *or*
1. export in Text file format (.txt). The text part will be saved in ASCII .txt format, the graph will be given a new file name, internally assigned in .jpg format *or*

Save the content of the Viewer and the output file through the Clipboard to store it in another program environment.

EXERCISE :

Use first EXCEL for the univariate regression and the National Data Set to relate :

Infant Mortality (the independent variable)

and

Fertilizer consumption

then use SPSS for a linear Regression with two explaining variables and add

CO² emissions

as an independent variable.



3. Multivariate and complex statistics

- ✓ If it comes to more complex statistical tasks, EXCEL has no direct solution (of course you can always enter a statistical formula yourself). Learn about more complex statistical procedures to respond to more specific and more complex questions concerning relations between variables

3.1. Cluster Analysis .

PRESENTATION:

International Data Set for environmental series

Problem:

find the resemblance between countries

BACKGROUND:

The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics. For example, you can group television shows into homogeneous categories based on viewer characteristics. This can be used to identify segments for marketing. Or you can group countries into homogeneous clusters so that comparable countries can be selected to test various political strategies.

Cluster analysis is a multivariate procedure for detecting groupings in the data. The objects in these groups may be cases or variables. A cluster analysis of cases resembles discriminant analysis in one respect—the researcher seeks to classify a set of objects into groups or categories, but, in cluster analysis, neither the number nor the members of the groups are known. That is, in cluster analysis, you begin with no knowledge of group membership and often do not know just how many clusters there are. A cluster analysis of variables resembles factor analysis because both procedures identify related groups of variables. However, factor analysis has an underlying theoretical model, while cluster analysis is more ad hoc. Clustering is a good technique to use in exploratory data analysis when you suspect the sample is not homogeneous.

SPSS provides two methods for clustering objects into categories: **Hierarchical Cluster Analysis** and **K-Means Cluster Analysis**. The former clusters either cases or variables; the latter, cases only. Each procedure has useful features.

Hierarchical clustering. In the hierarchical method, clustering begins by finding the closest pair of objects (cases or variables) according to a distance measure and combines them to form a cluster. The algorithm continues one step at a time, joining pairs of objects, pairs of clusters, or an object with a cluster, until all the data are in one cluster. The clustering steps are displayed in an icicle plot or tree (dendrogram). The method is *hierarchical* because once two objects or clusters are joined, they remain together until the final step. That is, a cluster formed in a later stage of the analysis contains clusters from an earlier stage which contain clusters from a still earlier stage.

K-means clustering. The K-Means Cluster Analysis procedure begins by using the values of the first k cases in the data file as temporary estimates of the k cluster means, where k is the number of clusters specified by the user. Initial cluster centers are formed by assigning each case in turn to the cluster with the closest center and then updating the center. Then, an iterative process is used to find the final cluster centers. At each step, cases are grouped into the cluster with the closest center, and the cluster centers are recomputed. This process continues until no further changes occur in the centers or until a maximum number of iterations is reached. You can specify cluster centers, and SPSS will allocate cases to your centers. This allows you to cluster new cases based on earlier results. The K-Means Cluster Analysis procedure is useful when you have a large number of cases.

3.2. Call the CLUSTER command.

Call the cluster command
[Analyze/Classify/Cluster]

If you are clustering cases, select at least one numeric variable. If you are clustering variables, select at least three numeric variables.

You may select a string variable to identify cases. Move the variable into Label Cases by.

The alternatives in the Cluster group allow you to form clusters either of cases or of variables. The controls in the Display group allow you to select Statistics and Plots. When these are selected, the corresponding Pushbuttons allow you to request additional statistics or plots.

Click on **Statistics** to request statistics.

Click on **Plots** to request plots.

Click on **Method** to determine the clustering method and the distance measure used in the analysis.

Click on **Save** to save cluster memberships as new variables. This option is not available if you have selected to perform the cluster analysis on variables.

Decisions to make before starting

In requesting your cluster analysis, you may want to consider the following points:

Standardization method. Variables with large values contribute more to the calculations of distance measures than those with small values. For example, a value of infant mortality could be 168 babies, while an increase in the population of a country might be 0.1%. One way to avoid this problem is to transform or re-express all variables to the same scale. For example, if you transform each variable to z scores, each new variable has a mean of 0 and a standard deviation of 1 (see the squared Euclidean distance measure discussion below, where distances are computed in original units and standardized units). Or you could put each variable on a range of 0 to 1, where the smallest value is 0 and the largest becomes 1.

Hierarchical Cluster Analysis provides several ways to standardize or transform the data in order to avoid problems caused by scale differences. With K-Means Cluster Analysis, you need to standardize the data before using the K-Means Cluster Analysis dialog box (for example, compute z scores in Descriptives).

Distance measure. The Hierarchical Cluster Analysis procedure provides approximately 37 distance or similarity measures for defining how different or alike two objects are. When two cases are very similar, the value of a distance measure is small and the value of a similarity measure is large. That is, *distances* measure how far apart two objects are and *similarities* measure how close they are. The squared Euclidean distance is used frequently as a distance measure for

clustering cases, and the usual Pearson correlation is used often for clustering variables. The **squared Euclidean distance** is the sum of the squared distances over all variables.

Uses of Cluster Analysis

As examples, consider using cluster analysis to group:
Countries, using health indicators such as the relative number of doctors, dentists, pharmacists, nurses, and hospital beds, the percentages of animal fat and starch in the diet, and life expectancy.

The Data

Cluster analysis can be used to analyze interval data, count data (frequencies), or binary data. It is important that variables are measured on comparable scales. As shown above, variables measured on larger scales contribute more than those measured on smaller scales, even if they empirically are less useful for classification. If your variables are measured on different scales, you can convert them to similar scales by standardizing them. The Hierarchical Cluster Analysis procedure provides a means to automatically standardize the variables in your analysis. When using the K-Means Cluster Analysis procedure, any standardization should be done prior to the cluster analysis.

Differences between K-Means and Hierarchical Cluster Analysis

The k-means clustering method handles large problems (200 or more cases) more easily. Hierarchical clustering computes a distance matrix with entries for every pair of cases (or variables), so large problems become unwieldy. More importantly, when the sample size is large, icicle plots and dendrograms are hard to read and interpret because they spread across many pages. For small data sets, icicle plots and dendrograms provide an excellent picture of just when each case (or variable) is joined with another, and distance matrices can also be informative.

By providing the distance from each case to its cluster center, k-means clustering characterizes whether or not a case is close to the others within its cluster or is an outlier. The size of *F* statistics in k-means' one-way ANOVA is useful for identifying variables that drive the clustering and also those that differ little across the clusters. In the k-means clustering method, by inputting cluster centers, you can classify new cases.

The K-Means Cluster Analysis procedure requires that you specify the number of clusters, so you may need to try several analyses (for example, requesting three, four, and five clusters). Alternatively, you might consider running a subset of the cases in the Hierarchical Cluster Analysis procedure to determine a reasonable number of clusters. Of course, with hierarchical clustering, you still have to make a judgment call about the number of clusters (by studying the graphical displays); there is no magical test that tells the number.

The Hierarchical Cluster Analysis procedure has many options for standardizing your data, computing distances, and linking clusters. With the K-Means Cluster Analysis procedure, you need to standardize your data before requesting the cluster analysis. The Euclidean distance metric is used automatically.

Hierarchical clustering excludes all cases with values missing for variables used in the analysis. K-means clustering has an Option that assigns cases to clusters based on distances computed from all variables with nonmissing values.

PRESENTATION:

International Data Set of 55 Series for 18 countries related to environmental topics
Use GDP and GNP indicators to cluster counties:

1. How many clusters
2. Which method to use
3. Use different presentation methods to follow the process of clustering

Output in SPSS

EXERCISE:

Use the International Data Set to Form Clusters using Series according to

- Health indicators,
- Life Expectancy,
- Environmental Indicators



Factor Analysis

Problem:

Find Indicators for what you can call Environmental quality, Environmental measures ,
Economical and Social well being

BACKGROUND:

Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors. For example, factor analysis can be used to explain the correlations in a battery of tests on the basis of factors that measure overall intelligence and mathematical and verbal skills. Or it can be used to determine the dimensions on which consumers rates coffees. These might be heartiness, body, and freshness.

Principal components and common factor analyses are often placed under the heading **Factor Analysis**. Although they are based on different mathematical models, they can be used on the same data, and both produce similar results. These procedures are often used in exploratory data analysis to:

- Study the correlations among a large number of interrelated quantitative variables by grouping the variables into a few factors; after grouping, the variables within each factor are more highly correlated with variables in that factor than with variables in other factors.
- Interpret each factor according to the meaning of the variables. For example, the answers to a set of six or seven questions that cluster together might measure the respondent's satisfaction with a product.
- Summarize many variables by a few factors. SPSS can compute a score for each factor that you can use as input variable(s) for t-tests, regression, analysis of variance, discriminant analysis, and so on.

Steps in a factor analysis. There are four main steps in factor analysis. They are introduced here, but the definition of new terms are found in the discussion and examples that follow.

1)The correlation or covariance matrix is computed. If a variable has very small correlations with all the others, you may consider eliminating it in the next run. Be sure, however, to check the size of its communality and loadings.

2)The factor loadings are estimated. Here, you decide whether the method of factor extraction is principal components or one of the factor analysis methods of extraction. We recommend beginning with principal components.

3)The loadings are rotated to make the loadings more interpretable. Rotation methods make the loadings for each factor either large or small, not in-between. After seeing these results, you may want to request fewer factors than chosen by default.

4)For each case, scores can be computed for each factor and saved for use as input variables in other procedures. You can also use saved scores to identify outliers and formulate a strategy for dealing with them.

Before you undertake these steps, be sure to screen each variable individually for outliers and skewed distributions. If you find problems, you may need to transform one or more variables. Also check how many values are missing. For the variables you are using, you may

want to compare descriptive statistics (means and standard deviations) for the sample that has no missing data (complete cases only) with descriptive statistics computed for each variable individually. If the sample size for cases with no values missing is considerably smaller than that for the total sample, compare listwise results with pairwise results at each step.

Call the FACTOR command

Call the Factor command
[Analyze/Data Reduction/Factor]

Select the variables for the factor analysis.

Click on **Descriptives** to request univariate statistics, a correlation matrix, or the initial (principal-components) solution.

Click on **Extraction** to specify the method of factor extraction and the criterion determining how many factors should be extracted.

Click on **Rotation** to specify a method for factor rotation.

Click on **Scores** to save factor scores as new variables or display the factor score coefficient matrix.

Click on **Options** to specify the display format of the factor loading and structure matrices.

.2. Variable names in SPSS

Principal components analysis (PCA).

Linear combinations of variables are useful for characterizing or accounting for the variation (spread) of each dimension in a multivariate space. Principal components analysis does this for you: the first linear combination of variables accounts for the largest amount of variation in the sample; the second for the next largest amount of variance in a dimension independent of the first, and so on. Successive components explain smaller and smaller portions of the total variance and are independent of one another.

In each solution, there are as many components as there are original variables. Ideally, for your data, the first few components should account for a large proportion of the variance of the original variables

.2. Variable names in SPSS

The variances of the components are commonly known as eigenvalues. The sizes of the eigenvalues describe the dispersion or shape of the cloud of data points **in a multivariate space that has one axis for each variable**. As an example of a three-dimensional space, imagine an American football with a pointed end placed in the corner of a box and tilted upward along 45 degree line. It is filled with tiny flies (the data points) frozen in space (the covering is clear so you can see inside). The flies fill the tilted ball and can be located in the 3-D space by using their length, width, and height coordinates measured from the corner of the box. This three-dimensional space has three principal components and each is a linear combination of the variables *length*, *width*, and *height*.

$a_1 \text{ length} + b_1 \text{ width} + c_1 \text{ height}$

$a_2 \text{ length} + b_2 \text{ width} + c_2 \text{ height}$

$a_3 \text{ length} + b_3 \text{ width} + c_3 \text{ height}$

where the coefficients a , b , and c are called **factor loadings**.

What if the flies are not spread through the three dimensional space evenly, but concentrate around an elliptical plate extending the length of the football? Then the third eigenvalue is considerably smaller than the second; and, if ignored, the first two components would account for most of the total variability of the three original variables— that is, two components summarize the variation measured for three variables.

For a useful factor analysis, there should be some strong correlations among the original variables.

Factor analysis. While components are linear combinations of the observed variables, factors are linear combinations of *unobserved* variables. The usual factor analysis model expresses each variable as a function of factors common to several variables and a factor unique to the variable:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \dots + a_{jm}F_m + U_j$$

where:

z_j = the j th standardized variable

F_i = the common factors

m = the number of factors *common* to all the variables

U_j = the factor unique to variable z_j

a_{ji} = the factor loadings

Ideally, the number of factors, m , will be small, and the contribution of the unique factors will also be small. The individual factor loadings, a_{ji} for each variable should be either very large or very small so each variable is associated with a minimal number of factors. Thus, you want to explain the observed correlations using as few factors as possible. The unique factors, U_j , are assumed to be uncorrelated with each other and with the common factors.

Rotation. Usually the initial factor extraction does not give interpretable factors. One of the purposes of **rotation** is to obtain factors that can be named and interpreted. That is, if you can make the large loadings larger than before and the smaller loadings smaller, then each variable is associated with a minimal number of factors. Hopefully, the variables that had strongly together on a particular factor will indicate a clear meaning with respect to the subject area at hand.



PRESENTATION:

National Set of 55 Series related to environmental topics

Use FACTOR to find out for the chosen country:

1. Are there any factors
2. How can they be called
3. Are there factors which can be related to
 - Environmental quality,
 - Environmental measures ,
 - Economical and Social well being



Output in SPSS

EXERCISE:

Use the National Set to find Factors using Series according to

- Environmental quality,
- Environmental measures ,
- Economical and Social well being

Calculate predictors with the regression method or factorial analysis:

How will develop :

Environmental quality

Environmental measures ,

Economical and Social well being

under which condition

How do these conditions differ for the found clusters

Calculate International factors

for Economical and Social well being

(are these factors transferable)

Group countries according to Found factors

(is this a valid statistical procedure)