Tutorial: R in the Statistical Office

NTTS 2015, Brussels

MATTHIAS TEMPL
Assoc.-Prof.
Vienna University of Technology & Statistics Austria &
Palacký University Olomouc & data-analysis OG
matthias.templ@gmail.com

VALENTIN TODOROV
Management Information Officer,
United Nations Industrial Development Organization (UNIDO)
v.todorov@unido.org

Course Objectives:

The open-source programming language and software environment R is nowadays the most widely used and the most popular software environment for statistics and data analysis. This tutorial provides an **overview** of important software packages used in official statistics and survey methodology and discusses the usefulness of R for daily work in a statistical office.

The *CRAN Task View on Official Statistics and Survey Methodology* is presented, which gives a comprehensive overview of those packages that proved to be useful in that area. Currently the task view is structured into the following topics: Complex Survey Design including point and variance estimation and calibration, editing, imputation, statistical disclosure control, seasonal adjustment, statistical record matching, small area estimation, seasonal adjustment, indices and indicators. UNIDO has published two research papers, authored by the lecturers, on the topic *R in the statistical office*. These papers present an overview of R, which focuses on the strengths of this statistical environment for the typical tasks performed in national and international statistical offices and outline some of the advantages of R using examples from the statistical production process of UNIDO where certain steps were either migrated or newly developed in R.

To demonstrate the usefulness of the considered packages for real-world problems, few specific applications will be presented, such as applications for generating publication quality graphics included in the UNIDO International Yearbook of Industrial Statistics; applications of specialized packages like VIM (visualization and imputation of missing values), simPopulation (generation of synthetic data), sdcMicro and sdcTable (both for statistical disclosure control), sparkTable (visualizing graphical tables in web sites and publication) or/and laeken (robust semi-parametric estimation of indicators). Several other practical questions are addressed, like data base connections, interfaces to other software, data formats and dynamical reporting facilities. Coming to the end of the tutorial we will find out that if someone searches for flexibility in data import/export, state-of-the-art methods, excellent features for presenting the data, high-performance computing, and if one really looks for an economical solution one inevitably arrives at R.

The Instructors

Matthias Templ: is associated professor at the Department of Statistics and Probability Theory at the Vienna University of Technology, researcher at the methods unit at Statistics Austria, consultant at Palacký University Olomouc and consultant and one of the directors and founders of

data-analysis OG. He is author and maintainer of the CRAN Task Views on „Official Statistics and Survey Methodology“ and he actively develop R in the area of Official Statistics and Computational Statistics for more than 10 years. He is main author of the `VIM` package for imputation and visualisation and the `sdcMicro` package and co-author of the `sdcMicroGUI`, `laeken`, `X12` and the `sparkTable` package and wrote several other packages in other fields of research. He is editor of the Austrian Journal of Statistics and associated editor for several other Journals. He published many papers in well-known journals in the area of Official Statistics.

Valentin Todorov (UNIDO): Valentin Todorov is a management information officer with the United Nations Industrial Development Organization (UNIDO). His main research activities include multivariate data analysis, computational statistics, robust statistics, official statistics and statistical information systems. Valentin Todorov has authored one book, book chapters and many scientific articles in the area of mathematical statistics and statistical information systems and acts as a reviewer for a number of international statistics journals. He has developed and maintains at CRAN several packages for the statistical software environment R.

Tentative Schedule

Morning session includes amongst other topics:

1. R as a mediator (data management and data exchange with R)
2. Handling of missing values and multiple imputation using `Amelia`
3. Automatic generation of reports (examples from the UNIDO's *International Yearbook of Industrial Statistics*)

Afternoon session may include amongst other topics:

1. Data screening and data validation with R
2. Overview of some useful packages (`X12`, `editrules`, `tableplot`, `survey`)
3. Aspects of statistical disclosure control using `sdcMicro`
4. Estimation of social inclusion indicators using `laeken`
5. Visualisation of indicators using `sparkTable`

Prerequisites

This is a BYOD (bring your own device) tutorial. It is recommended to install the latest version of R (`http://cran.r-project.org`) and the following by typing into R:

```
install.packages("ctv")
require(ctv)
install.views("OfficialStatistics")
```

As a script editor it is recommended to install RStudio (`http://www.rstudio.com`.

Ideally the participants have basic knowledge in R. In addition, basic knowledge in statistics, especially in official statistics, are desirable. Without any knowledge on these topics it will be hardly possible to follow details presented in the course, but to gain a general overview might be possible.