

The Software Environment R for Official Statistics and Survey Methodology

Matthias Templ
TU WIEN &
Statistics Austria

Valentin Todorov
UNIDO

Abstract

The open-source programming language and software environment R is currently one of the most widely used and popular software tools for statistics and data analysis. This contribution provides an overview of important R packages used in official statistics and survey methodology and discusses the usefulness of R in the daily work of a statistical office. Examples of activities and developments in R related projects in several national and international statistical offices are given. The focus is not only on the internal infrastructure that national and international statistical offices provide for using R but also on some interesting R related projects carried out in those institutes. Two particular packages (laeken and sdcMicro) and one data set (Statistics on Earnings Survey) are used to illustrate the usefulness (and the user-friendliness) of R and to present methods available in R. In addition, the access to international statistical databases like WDI of World Bank and UN COMTRADE with R is illustrated.

Keywords: official statistics, survey methodology, R.

1. Some general statements on R

The R Core Team (<http://www.r-project.org/>) defines R as an environment rather than a statistical system or programming language. R is an integrated suite of software facilities for data manipulation, calculation and graphical display, which includes:

- a suite of operators for calculations on arrays, mostly written in C and integrated in R;
- comprehensive, coherent and integrated collection of methods for data analysis;
- can be extended with (add-on) packages;
- graphical facilities for data analysis and display, either on-screen or in hard copy. It features trellis graphics (Sarkar 2008) and an implementation of the grammar of graphics book (Wilkinson and Wills 2005; Wickham 2009);
- a well-developed, simple and effective programming language which includes conditional statements, loops, user-defined recursive functions and input and output facilities;
- a flexible object-oriented system facilitating code reuse;

- high performance computing with interfaces to compiled code and facilities for parallel and grid computing;
- an environment that allows communication with many other software tools.

Each R package provides a structured standard documentation including code application examples. Further documents (so called vignettes [Leisch 2003](#)) are also available showing more applications of the packages and illustrating dependencies between the implemented functions and methods.

R has become an essential tool for statistics and data science ([Godfrey 2013](#)). Also companies in the area of social media (Google, Facebook, Twitter, Mozilla Corporation), in the banking world (Bank of America, ANZ Bank, Simple), in the food and pharmaceutical area (FDA, Merck, Pfizer), finance (Lloyd, London, Thomas Cook), technology companies (e.g. Revolution Analytics as subsidiary of Microsoft), car construction and logistic companies (Ford, John Deere, Uber), newspapers (The New York Times, New Scientist), and companies in many other areas use R in a professional context (see also [Gentlemen 2009](#); [Tippmann 2015](#)).

Outline of the paper

This contribution focuses on the use of R in national and international statistical organizations, giving the readers an overview of the usefulness of R in the area of official statistics. In Section 2 we first give some impression about systems and software used in daily work at the statistical office. This shows which variety of software is used in the daily work and how R could serve as a mediator between and within these software products.

Survey methodology/statistics is one large part of official statistics. Typically, the data in this area are of complex nature, having hierarchical structures and sampled from finite populations using complex sampling designs. Due to non-response, measurement errors and additional available information, many of the methods are focused on pre-processing of the data while the main part of other methods are related to data summarization and dissemination. A summary of available tools in different topics in survey methodology is therefore given in Section 3, which enhances the CRAN Task View on Official Statistics that is maintained by the authors of this paper.

Having presented a list of useful packages, the use of these packages and the strategy of using R in national statistical offices as well as in international statistical offices are presented in Section 4. This should give the readers an impression of the current status of the usage of R and ideas how to integrate R into the production system. However, it is out of scope to give a complete picture of the usage of R in each institution. Therefore a sample of selected offices is taken and reported.

Two examples of applying R on real data from official statistics cannot show all the useful packages and functions in the area of official statistics, but they show how efficiently R can be used for a specific task. In the first example of Section 5, the estimation of indicators is in focus while the second example shows the application of statistical disclosure control methods. Both topics are of main interest in official statistics. In Section 5.3 the access to international databases is shown. For the access to the world development indicators an R package is used while JavaScript object notation (JSON) format is used to access data from UN COMTRADE.

2. R in the statistical office

National and international statistical offices' main responsibility is to collect and publish empirical information about our society and economy, which become an important economic and social factor. The huge information requirements of our society have led to the development of sophisticated methods to collect, process, analyze and supply information. It is the role of national statistical offices to provide reliably collected and expertly analyzed political, social

and economic information. This information provides a basis for political decision-making; Its application and use for this purpose has become of increasing importance for the general public.

For data processing and data analysis in national or international statistical organization, several well-established statistical software packages are often available (see also [Todorov and Templ 2012](#)):

- (i) SAS[®] because of its traditional position in these organizations (if the necessary funds are available), its ability to handle large data sets and its availability on any platform, including mainframes;
- (ii) SPSS[®] is considered user friendly because of its point-and-click interface (albeit still providing the so-called syntax);
- (iii) Stata[®] is likely the most cost-effective among the three and, in terms of its design, is particularly suitable for handling data generated from complex survey designs (as is the case in many NSOs).

However, if the objective is

- flexibility in reading, manipulating and writing data,
- availability of recent statistical methodology,
- versatile presentation capabilities for generating tables and graphics which can readily be used in text processing systems such as L^AT_EX (or Microsoft Word),
- creating dynamical reports using, e.g., **rmarkdown** ([Allaire, McPherson, Xie, Wickham, Cheng, and Allen 2014](#)), **Sweave** ([Leisch and Rossini 2003](#)), **brew** ([Horner 2011](#)) or the **knitr** ([Xie 2013](#)) package,
- to build web-based applications using **shiny** ([RStudio Inc. 2014](#)), and last but not least,
- a particularly economical solution,

R is the answer (see also [Todorov and Templ 2012](#)). The integration of all types of modern tools for scientific computing, programming and management of data and files into one environment is possible. Such an environment combines the capacities of R with editors that allow syntax highlighting and code completion, the use of modern version control systems for code and file management or a modern document markup languages such as L^AT_EX or Mark-down, or interfaces to general purpose programming languages such as C, C++ or Java as well as easy-to-use (automatic) connections to powerful workstations. There exist editors that provide a complete programming environment for R. For example, eclipse with the extension STATET, an eclipse interface for R (see <http://www.walware.de/goto/statet>) or the modified eclipse IDE from *Open Analytics* called Architect (<http://www.openanalytics.eu/architect>), provide not only syntax highlighting, a defined project philosophy and interaction with R, but also integrate C++, Java, L^AT_EX, **Sweave**, Subversion, server-connection facilities and many more. A very popular IDE for R nowadays is RStudio (see <http://support.rstudio.com>), which includes these features and additionally includes an integration of the packages **shiny** and **rmarkdown** (and related tools for creating slides in HTML – *RPresentation*), i.e. it provides a modern scientific computing environment, well designed and easy to use.

Although R has powerful statistical analysis features, data manipulation tools and versatile presentation capabilities, it has not yet become a standard statistical package in national and international statistical offices. This is mainly attributable to the widespread opinion that R is difficult to learn and has a very steep learning curve, which is true for learners without any programming skills. However, GUI's which display the underlying produced code are available as well ([Fox 2005](#)), and interfaces to all popular GUI toolkits are available.

In the daily work of a national or international statistical organization, for example, at Statistics Austria or at UNIDO, different systems for data analysis and data processing are used. Data exchange between statistical systems (like SAS[®], SPSS[®], EViews, Stata[®], Microsoft Excel), database systems (like Microsoft Access, MySQL, IBM DB2) or output formats (like HTML, XML) is often required. Also note that the importance of statistical data and metadata exchange format (like SDMX) is continuously growing. In this respect, R offers very flexible import and export interfaces either through its base installation or through add-on packages which are available from CRAN. For example, the packages **XML** (Temple Lang 2013) and **xml2** (Wickham 2015b) allow to read XML files. For importing delimited files, fixed width files and web log files it is worth mentioning the package **readr** (Wickham and Francois 2015) which is supposed to be faster than the available functions in base R. See also the **fread** function from package **data.table** (Dowle, Short, Lianoglou, and Srinivasan 2014) that is also substantially faster than **read.csv** from base R. The packages **XLConnect** (Mirai Solutions GmbH 2015) and **readxl** (Wickham 2015a) import Microsoft Excel files (.xls and .xlsx) into R. The packages **foreign** (R Core Team 2015) and a newer promising package called **haven** (Wickham and Miller 2015) allow to read SPSS[®], Stata[®] and SAS[®] files from within R.

The connection to all major database systems is easily established with the packages **ROracle**, **RMySQL**, **RSQLite**, **RmSQL**, **RPgSQL**, **RODBC** and **RJDBC**. The **DBI** package provides an extra abstraction layer, used by the former mentioned packages. The integration of other statistical systems is made possible through packages like **RWeka**, **X12** and **RExcel**, while data exchange among these systems is facilitated by the package **foreign**. For more specific applications like web page generation or data and metadata exchange between organizations, the packages **R2HTML**, **sdmxer** or **RSDMX** can be used.

Data manipulation – in general but in any case with large data – can be best done with the package **dplyr** (Wickham and Francois 2014) or the **data.table** (Dowle *et al.* 2014) package. Functions for data manipulation in **dplyr** are implemented in C++ and thus the computational speed is often much faster than the data manipulation functions of the base packages. The syntax of **dplyr** is easy to learn and it is possible to write **dplyr** syntax in *data pipelines* that is internally provided by package **magrittr** (Bache and Wickham 2014). **data.table** (consisting of C code) even slightly outperforms **dplyr** but the syntax is not easy to learn, since especially indexing is done differently from the base packages.

In the case of large data files which exceed available RAM, interfaces to (relational) database management systems might be useful. The large data set problem can also be resolved by using either the **filehash**, **LaF**, **ff**, **ffbase** or **bigmemory** packages or by connecting to powerful workstations with **Rserve** (see <http://www.rforge.net/Rserve/>). Note that the RStudio server enables you to provide a browser based interface (the RStudio IDE) to a version of R running on a remote Linux server. This has some advantages such as to access your R workspace from any location, to allow multiple users to access powerful hardware and having installation of R packages and related features centralized.

R is designed to allow for parallel computing using multiple cores or CPU's and the recommended package for this task is the R package **parallel**. Several other packages can also be used for parallel computing, such as **snow** or **foreach**. Note that for medium or large data sets, parallel computing can be very slow on Microsoft Windows machines since the *fork* system call is only available on POSIX compliant platforms (e.g. Unix based operating systems). There exist user friendly interfaces for integrating compiled code (for example, the package **Rcpp**, **inline** and **rJava**) and features for code profiling (the package **profr** and **proftools**). Worth to mention is also an integration of scripting languages such as JavaScript, see for example (amongst others) package **shinyjs** (Attali 2016). In addition, R offers interfaces to almost all commercial and open-source linear program solvers which are often needed for special tasks in survey methodology.

3. R for survey statistics

R includes several methods that are helpful for data processing and survey methodology in statistical offices and organizations which usually deal with complex data sets from finite populations. The CRAN task view on *Official Statistics and Survey Methodology* (<http://cran.r-project.org/view=OfficialStatistics>) contains a list of packages which include methods typically used in official statistics and survey methodology. Below we list those packages and briefly outline their functionalities.

3.1. Complex survey designs

The package **sampling** includes various algorithms (Brewer, Midzuno, pps, systematic, Sampford, balanced (cluster or stratified) sampling via the cube method, etc.) for drawing survey samples, as well as functionality to calibrate the design weights (see, e.g., Tillé 2006).

For estimation purposes and to work with already drawn survey samples, the package **survey** (Lumley 2010) – the standard package for that task – can be used once the given survey design has been specified (stratified sampling design, cluster sampling, multi-stage sampling and pps sampling with or without replacement). The resulting object can be used to estimate (Horvitz-Thompson-) totals, means, ratios and quantiles for domains or the entire survey sample, and to apply regression models. Variance estimation for means, totals and ratios can either be done by Taylor linearization or resampling (BRR, jackknife, bootstrap or user-defined).

As an add-on package, **ReGenesees** uses facilities from and extends the **survey** package (not on CRAN, see <http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/regeneeses>). This package also includes a GUI.

The package **EVER** (Estimation of Variance by Efficient Replication) provides variance estimation for complex designs by delete-a-group jackknife replication for (Horvitz-Thompson-) totals, means, absolute and relative frequency distributions, contingency tables, ratios, quantiles and regression coefficients, even for domains.

The **laeken** package (Alfons and Templ 2013) implements functions to estimate certain social inclusion indicators (at-risk-of-poverty rate, quintile share ratio, relative median risk-of-poverty gap, Gini coefficient), including their variance for domains and strata based on (calibrated) bootstrap resampling.

To perform simulation studies in official statistics, the package **simFrame** (Alfons, Templ, and Filzmoser 2010) provides a framework for comparing different point and variance estimators under different survey designs as well different scenarios for missing values, representative and non-representative outliers.

Other packages are available for selecting samples with specific designs: **pps**, **sampling** and **SamplingStrata**.

3.2. Calibration

To calibrate the sampling weights to precisely match the population characteristics and/or to calibrate for unit non-responses, the package **survey** allows for post-stratification, generalized raking/calibration, generalized regression (GREG) estimation and trimming of weights.

The **EVER** package includes facilities (function `kottcalibrate()`) for calibrating on known population characteristics, on marginal distributions or joint distributions of categorical variables, or on totals of quantitative variables.

The `calib()` function in the package **sampling** allows to calibrate for non-response (with response homogeneity groups) for stratified samples. The implementation in **laeken** (and package **simPop**) (function `calibWeights()`) is similar but possibly faster.



3.3. Editing and visual inspection of microdata

The package **editrules** (de Jonge and van der Loo 2012) provides tools for editing and error localization using the Fellegi-Holt principle and categorical constraints. It converts readable linear (in)equalities into matrix form, which can then be applied for editing the given data set. The package **deducorrect** depends on the package **editrules** and applies deductive correction of simple rounding, typing and sign errors based on balanced edits. Values are changed so that the given balanced edits are complete.

The package **rrcovNA** (Todorov, Templ, and Filzmoser 2011) provides robust location and scatter estimation and robust principal component analysis with a high breakdown point for incomplete data. It can thus be used to find representative and non-representative outliers.

The package **VIM** (Templ, Alfons, and Filzmoser 2012) can be used for visual inspection of microdata. It is possible to visualize missing values using suitable plot methods and to analyze missing values structure in microdata using univariate, bivariate, multiple and multivariate plots. The information on missing values from specified variables is highlighted in selected variables. **VIM** can also evaluate imputations visually. Moreover, the package **VIMGUI** (Schopfhause, Templ, Alfons, Kowarik, and Prantner 2014) provides a point and click graphical user interface (GUI).

The package **tabplot** (Tennekes, de Jonge, and Daas 2013) entails the table plot visualization method, which is used to profile or explore large statistical data sets. Up to a dozen variables are shown column-wise as bar charts (numeric variables) or stacked bar charts (factors). Hierarchies can be visualized with the package **treemap**.

Visual analysis of data is important to understand the main characteristics, main trends and relationships in data sets and it can be used to assess the data quality. Using the R package **sparkTable** (Kowarik, Meindl, and Templ 2014a), statistical tables holding quantitative information can be enhanced by including spark-type graphs such as sparklines  and sparkbars . These kind of graphics have initially been proposed by Tufte (2001) and are considered as simple, intense and illustrative graphs that are small enough to fit in a single line. Thus, they can easily enrich tables and texts with additional information in a comprehensive visual way. The R-package **sparkTable** uses a clean **S4**-class design and provides methods to create different types of sparkgraphs that can be used in webpages, presentations and text documents. With the GUI, graphical parameters can be interactively changed, variables can be sorted, and graphs can be added/removed in an interactive manner. Thereby it is possible to produce custom-tailored graphical tables – standard tables that are enriched with graphs – that can be displayed in a browser and exported to various formats.

3.4. Imputation

A distinction is made between interactive model-based methods, k -nearest neighbor methods (k NN) and miscellaneous methods. However, the criteria for using a given method depend on the scale of the data, which in official statistics are typically a mixture of continuous, semi-continuous, binary, categorical and count variables. Note that only few imputation methods can deal with mixed types of variables and semi-continuous ones, and only the methods in the package **VIM** account for robustness issues.

Expectation-Maximization (EM) based imputation methods are offered by the packages **mi** (Yu-Sung, Gelman, Hill, and Yajima 2011), **mice** (van Buuren and Groothuis-Oudshoorn 2011), **Amelia** (Honaker, King, and Blackwell 2011), **VIM** and **mix** (Schafer 1997). The package **mi** provides the iterative EM-based multiple Bayesian regression imputation of missing values and checking of the regression models used, whereas the regression models for each variable can also be defined by the user. The data set may consist of continuous, semi-continuous, binary, categorical and/or count variables. The package **mice** (van Buuren and Groothuis-Oudshoorn 2011) provides iterative EM-based multiple regression imputation as well, and the data set may consist of continuous, binary, categorical and/or count variables.

Multiple imputation in which first bootstrap samples are drawn for EM-based imputation can be carried out with **Amelia** (Honaker *et al.* 2011). It is also possible to impute longitudinal data. The package **VIM** offers EM-based multiple imputation (function `irmi()`) using robust estimations (Templ, Kowarik, and Filzmoser 2011), which adequately deal with data including outliers. It can handle data consisting of continuous, semi-continuous, binary, categorical and/or count variables.

Nearest neighbor (NN) imputation methods are also included in the **VIM** package. It provides implementation of the popular sequential and random (within a domain) hot-deck algorithm, and also a fast k NN algorithm which can be performed for large data sets. It uses a modification of the *Gower Distance* to deal with a mixture of numerical, categorical, ordered, continuous and semi-continuous variables.

3.5. Statistical disclosure control

Data from statistical agencies and other institutions are often confidential in its raw form. One of the main tasks of data providers is to modify the original data in order to guarantee that no statistical unit can be re-identified and, at the same time, to minimize the loss of information. For microdata perturbation, the package **sdcMicro** (Templ, Meindl, Kowarik, and Chen 2014c; Templ, Kowarik, and Meindl 2015) can be used to generate confidential (micro)data, i.e. to generate public- and scientific-use files. All methods are implemented in C++ to allow fast computations. The package **sdcMicroGUI** (Templ, Meindl, and Kowarik 2014b) also provides a GUI (Kowarik, Templ, Meindl, and Fontenau 2014c; Templ *et al.* 2014b).

To simulate synthetic data, the package **simPop** (Alfons, Kraft, Templ, and Filzmoser 2011b; Meindl, Templ, Alfons, and Kowarik 2014) offers methods for the simulation of synthetic, confidential, close-to-reality populations for surveys based on sample data. Such population data can then be used for extensive simulation studies in official statistics using, for example, the package **simFrame** (Alfons *et al.* 2010). Another package, **synthpop**, offers also methods to simulate populations, but it is not designed for dealing with hierarchical structures of the data.

For tabular data, the package **sdcTable** (Templ and Meindl 2010) can be used to provide confidential (hierarchical) tabular data. It includes techniques to solve the secondary cell suppression problem. A method is included that protects the complete hierarchical, multi-dimensional table at once, and therefore it is only suitable for small problems. The package also offers interfaces to various commercial and open-source linear program solvers.

3.6. Time series analysis and seasonal adjustment

For a general time series methodology, we refer to the CRAN Task View *TimeSeries* on CRAN. Specifically for survey methodology, the decomposition of time series can be done using the function `decompose()`. For a more advanced decomposition the functions `st1()` and `StructTS()` can be used. All these functions are available in the base **stats** package. Many powerful tools for seasonal adjustment can be accessed via the R package **x12** and **x12gui** (Kowarik, Meraner, Templ, and Schopfhauser 2014b). It provides a wrapper function and GUI for the **X12 binaries**, which have to be installed first. Another package, available on CRAN is **seasonal**, which supports also *SEATS Specification Files* from TRAMO-SEATS, a software for seasonal adjustment developed by the Bank of Spain.

3.7. Statistical matching and record linkage

The package **StatMatch** (D’Orazio, Di Zio, and Scanu 2006) provides functions for conducting statistical matching between two data sources sharing a number of common variables. It creates a synthetic data set after matching of two data sources via a likelihood approach or hot-deck. **MatchIt** allows nearest neighbor matching, exact matching, optimal matching and

full matching, among other matching methods.

The package **RecordLinkage** (Borg and Sariyar 2015) provides functions for linking and de-duplicating data sets. It can be used to perform and evaluate different record linkage methods. A stochastic framework is implemented which calculates weights through an EM algorithm. Machine learning methods are utilized, including decision trees (package **rpart**), adaboost (package **ada**), neural nets (package **nnet**) and support vector machines (see package **e1071** and package **kernlab**). The generation of record pairs and comparison patterns from single data items are provided as well. Comparison patterns can be chosen to be binary or based on some string metrics. In order to reduce computation time and memory usage, blocking can be used.

It is worth mentioning the package **stringdist** (van der Loo 2014), which implements various methods for string comparison.

3.8. Small area estimation

The package **hbsae** (Boonstra 2012) provides functions to compute small area estimates based on a basic area or unit-level model. The model is fit using restricted maximum likelihood (REML), or in a hierarchical Bayesian way. Auxiliary information can be either counts resulting from categorical variables or means from continuous population information.

The package **rsae** (Schoch 2014) provides functions to estimate the parameters of the basic unit-level small area estimation (SAE) model (aka nested error regression model) by means of ML or robust ML. On the basis of the estimated parameters, robust predictions of the area-specific means are computed (incl. MSE estimates; parametric bootstrap). However, the current version (rsae 0.1-5) does not allow for categorical independent variables.

The package **nlme** provides facilities to fit Gaussian linear and nonlinear mixed-effects models and **lme4** includes facilities to fit linear and generalized linear mixed-effects model, both used in small area estimation. With package **JoSAE** (Breidenbach 2013), point and variance estimation for the generalized regression (GREG) and a unit level empirical best linear unbiased prediction (EBLUP) estimators can be made at domain level. It basically provides wrapper functions to the **lme** package that is used to fit the basic random effects models. Package **saeSim** (Warnholz and Schmid 2015) provides tools for simulation of synthetic data for small area related tasks.

3.9. Indices and indicators

A comprehensive collection of indicator methodology is included in the package **laeken** which helps to estimate popular risk-of-poverty and inequality indicators (at-risk-of-poverty rate, quintile share ratio, relative median risk-of-poverty gap, Gini coefficient, gender pay gap). In addition, standard and robust methods for tail modeling of Pareto distributions are provided for the semi-parametric estimation of indicators from continuous univariate distributions such as income variables. Classes for the resulting indicators are defined as well as print, summary, plotting and subsetting methods for objects of these classes. The variances of the indicators can be estimated via a calibrated bootstrap approach (Alfons and Templ 2013).

Various indicators of poverty, concentration and inequality are included in the package **ineq** (Zeileis 2014). It provides some basic tools like Lorenz curves and Pen's parade graph. However, it is not designed to deal with sampling weights directly (these could only be approximately emulated via `rep(x, weights)`). The package **IC2** includes three inequality indices: extended Gini, Atkinson and Generalized Entropy. It can deal with sampling weights and subgroup decomposition is also supported.

The function `priceIndex()` from the package **micEcon** can be used to calculate price indices. For visualization purposes, the package **sparkTable** offers tools to produce scalable sparklines for webpages and reports, as already mentioned in Section 3.3. It also contains visualization tools for presenting indicators in checker plots—a grid-based representation of possible com-

View on Official Statistics and Survey Methodology. Figure 3 also presents the download statistics from RStudio's repository and the most popular packages listed on the CRAN Task View on Official Statistics and Survey Methodology (note that a similar analysis was carried out at <http://www.r-bloggers.com/finally-tracking-cran-packages-downloads/>). The figures underestimate the download by not considering all CRAN mirrors and on the other hand overestimation is caused by counting also package updates. The packages **Hmisc**, **nlme** and **lme4** are downloaded with highest frequency. Note that these packages are also used for tasks not related to the official statistics and survey methodology. Therefore, the most widely used packages for survey tasks might be **survey** and **mice** (more than 250 times per week on average). In any case, the download peak refers to packages that include imputation methods (**mice**, **mitools**, **Amelia**, **VIM**, **mi**), which suggests the extensive use of this methodology.

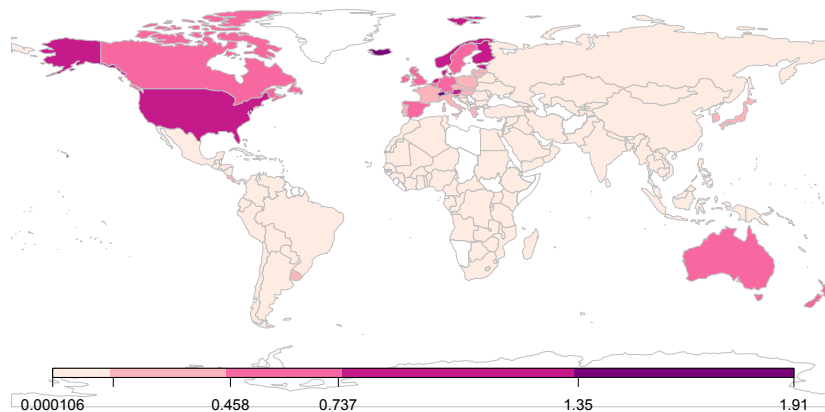


Figure 2: Choropleth map showing the downloads of R packages included in the CRAN Task View on Official Statistics and Survey Methodology over the world. The number of downloads are presented per capita (i.e. normalized by dividing by the population counts).

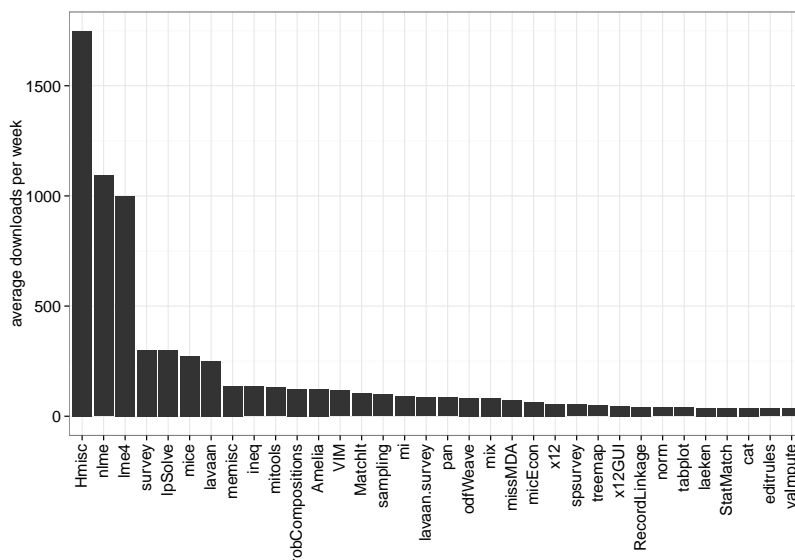


Figure 3: Average number of downloads per week (from October 2012 till March 2014) for the most downloaded packages listed on the CRAN Task View on Official Statistics and Survey Methodology.

4. The use of R in selected national statistical offices

Statistical organizations need a strategy for using and distributing software to their employees. The national statistical offices of Canada, Austria, Netherlands, Italy, USA, UK, the statistical offices of a few other countries and several international agencies are quite active in using R. This is recognized by the number of contributions at international conferences, see for example the official statistics sessions at the useR! 2013 conference in Albacete, the R session at the International Conference on Establishment Surveys in Montreal 2012, as well as the UNECE data editing work sessions in 2012, 2013 and 2014. The latter had sessions focused on the use of R in data editing and imputation. Tutorials at the NTTS 2015 and the Q2014 on the use of R in the statistical office were presented.

Moving to R seems to be slow due to the lack of strong programming knowledge and availability of many legacy code developed in other statistical software environments. However, new and innovative projects using R are carried out, some of them are listed in the following subsections. Examples will be given of how R is used in selected national statistical offices.

4.1. Use of R at Statistics Austria

At Statistics Austria, R is currently installed on more than 65 computers and on virtual servers. This is useful for tasks involving large memory requirements or to put content on the web, e.g. via **shiny** (RStudio Inc. 2014).

The leading R-team at Statistics Austria consists of three methods division experts. In addition, each department has nominated one person as first contact in case of questions and problems. Furthermore, the following organizational setup has been chosen:

- the R-experts at the methods unit (the administrators) take care that the version used is always up-to-date and they also decide on a GUI front-end which packages should be installed with the default installation. They place all necessary files (R, RStudio, packages, documentation, examples) on a particular server, in the following termed *container*;
- the IT department takes these files and deploys the R-installation including the front-end (RStudio) to users. This ensures that only one standardized software package is installed on all computers;
- the general R-support is centralized through a mailing list (apart from direct questions that can be answered by first-contact persons);
- an internal wiki was created and is used to collect know-how;
- the administrators define – together with the IT department – access rights for users on the servers, the mailing list, wiki, file depot, etc. In this situation, basically two user groups exist:
 1. R-administrators: have read and write access and are responsible for the folder containing all software package, documentation, wiki, etc. Additionally, administrators have full access to the mailing list “R-Support”.
 2. R-team: team-members have read-only access to the R container and read and write access to the wiki.

Currently five R courses are offered for the employees at Statistics Austria. The introductory course consists of 12 hours tutorial (3×4 hours), the aim is to reach a certain level of knowledge for all participants. The target group is constituted by beginners and regular R users who learned R in self-study. For the latter group many fundamental insights to the software are presented which are mostly new even for experienced users. The course covers topics on data types, data import/export (including database connections), syntax, data manipulation

(including presentation of important add-on packages such as **dplyr** and **data.table**) and basic object-orientation features. Ex-cathedra teaching is followed by exercises and R sessions in which the trainers interactively give additional insights.

The advanced training course consists of four-module courses between 4 and 8 hours each. This course covers topics on graphics (**graphics**, **grid**, **ggplot2**, **ggmap**, **ggvis**), classes and object-orientation (**S3** classes as well as a brief introduction to **S4** classes), dynamic reporting (**Sweave**, **knitr**, **brew**, **rmarkdown**), R development issues (profiling, debugging, benchmarking, basic packaging), web-applications (**shiny**), data manipulation (**dplyr**, **data.table**) and survey statistics using **survey**. The courses give an overview of other useful packages for key-tasks in official statistics.

At Statistics Austria, methodological training is offered to staff. In these courses, R is intensively used for teaching purposes, but participants do not get in direct touch with R. The reason was that for these courses no requirements with respect to any software or programming language should be set. Thus, a blended learning feedback system was developed (Dinges, Kowarik, Meindl, and Templ 2011). At the start of a course, participants fill out an online questionnaire. The collected data are then automatically used in the exercises and are also incorporated into the presentation slides. Participants are able to identify their data in various graphics, tables and other output. At various times, participants have to do interactive exercises using point and click directly in the browser. An online client/server based tool was developed which includes (among others) single- and multiple choice questions, animated and interactive examples. All clicks and answers from the participants are automatically saved on the server and aggregated statistics (feedback) are generated automatically in order to show clearly if the examples were correctly solved. In a teacher-interface, trainers can select certain examples which are then available for course participants. They also get feedback about how many participants have already solved which question and the correctness of the solutions. In the student-interface, the current exercises are available and listed ready to be selected.

Beside the development of the packages **laeken**, **sparkTable**, **sdcTable**, **sdcMicro**, **sdcMicroGUI**, **VIM**, **VIMGUI**, **x12** and **x12GUI**, R is applied in the production process for sampling, editing and imputation, estimation, analysis and output generation for several surveys. One example could be the estimation regarding PEAAC (Program for the International Assessment of Adult Competencies) or even the tourism statistics as well as the higher education forecast where R is used for everything, also for producing reports (see, e.g., Radinger, Nachtmann, Peterbauer, Reif, Hanika, Kowarik, and Lehner 2014).

4.2. Use of R at the national statistical office at UK

The National Statistical Office of UK started to use R version 2.0.1. in 2004. Employees training is done since 2005. In 2011 an R testing group was established consisting of members of the methods and IT department, with the aim of testing disclosure control tools and standard graphics using R. Since 2012 an R development group has been installed, whose objective was to test if R, and related specialized packages, are ready for use in the production environment. Several applications were investigated in the past. R was used to call **X12ARIMA**. In addition, the **dlnm** package (Petris 2010) was used to calculate the unemployment statistics. The **survey** package is used for the Labor Force Survey (panel design). The survey error correlation problem induced by this rotating panel is considered, specifying a multivariate model according to time series of five waves (a selected household will be surveyed five times in row, 1/5 of households are exchanged) of the rotating panel. The **spatstat** package and its kernel smoothing features (Baddeley and Turner 2005) are used to visualize crime data at postcode level. The **MortalitySmooth** package (Camarda 2012) has been used for mortality rates estimation (Brown, Mills, Ayoubkhani, and Gallop 2013). Hereby, the (smoothed) mortality against age per year are presented in heat maps.

Currently, the **survey** and **ReGenesees** packages are tested and other statistical functions are implemented for the national accounts database *CORD*.

4.3. Use of R at the national statistical institute of Romania

The National Statistical Institute of Romania established the Romanian R team in 2013. They organized already two international workshops on *New Challenges for Statistical Software - The Use of R in Official Statistics*, the last held in Bucharest in April 2015.

One course was given with the title *Introduction in SAE estimation techniques with application in R* whereas the **JoSAE** package was used. Several other courses on R are planned.

Various R packages play an important role for the Business Register Department, due to the need of using of administrative data in the production of business statistics. Hereby, R is used because of its flexibility and powerful tools for editing, validation of data and data imputation. The quality of data is in general evaluated and reported using existing tools in R. Probabilistic record linkage methods are applied to merge enterprises from different data sources on the basis of names and address information. The Department of Indicators on Population and International Migration uses the package **JoSAE** for the application of small area estimation techniques. It is used to produce data on annual international migrant stocks. The Department of Social Statistics tests the use of R for sampling. Particularly packages **ReGenesees**, **vardpoor** and **survey** are tested for use in the production according to variance estimation for household surveys.

More information can be found in (Dobre and Adam 2014).

4.4. Use of R at the national statistical institute of Serbia

The Statistical Office of Serbia uses the **laeken** package for poverty estimation. They compared the results obtained with **laeken** to those obtained with SAS[®] Macros provided by Eurostat. They also use R for estimations on the monthly retail trade survey. They use the package **XLConnect** to read and write Microsoft Excel files from within R.

4.5. Use of R at Statistics Netherlands

Statistics Netherlands started using R in a systematic manner already in 2010, see van der Loo (2012). A knowledge center was built up, an internal wiki provides code examples and serves as a platform for knowledge sharing. Every employee who wants to use R participates in training courses that are offered internally in the office. Statistics Netherlands distinguishes between three installations of R:

- the production installation is the smallest one,
- the analysts installation includes more packages and
- the researchers installation includes all tools which are useful for the development of R code including the *RTools* (<https://cran.r-project.org/bin/windows/Rtools/>) which facilitate the infrastructure for R package building on Microsoft Windows platforms.

Currently R is installed on approximately 160 individual computers and it is used mainly for data manipulation tasks, regression analysis, visualizations and data editing. Examples of R use in the statistical production process include the estimation of the Dutch Hospital Standardized Mortality Ratio (HSMR), the estimation of certain unemployment figures, estimation of tourist accommodations, and manipulation of supply and demand tables for National Accounts, see van der Loo (2012). Statistics Netherlands uses R also for data collection and web crawling with web robots, and for collecting data for compilation of price statistics.

Several packages were developed by Statistics Netherlands. The **editrules** package for data editing and the **deducorrect** package for deductive correction and deductive imputation, see Section 3.3. With the **rspa** package numerical records can be modified to satisfy edit rules.

These packages are used for automatic data editing system in child care center statistics. .NET/SQL is used to communicate between R and the database.

Also packages for visualization were developed by Statistics Netherlands and submitted to CRAN: **treemap**, **tableplot** and **tableplot3**. The package **LaF** can import large ASCII files. In addition, Statistics Netherlands also developed the packages **stringdist**, **docopt**, **ffbase**, **daff** and **whisker**.

4.6. Use of R at UNIDO

The statistical business process of international organizations is slightly different from that in the national statistical offices. International organizations like the United Nations Industrial Development Organisation (UNIDO) do not carry out surveys, but collect and aggregate data from national authorities (statistical offices, ministries, governmental departments) and create multivariate cross-sectional time series for their analysis and for further dissemination. The statistical activities of UNIDO are defined by its responsibility to provide the international community with global industrial statistics and meet internal data requirements to support the development and research program of the organization. Currently, UNIDO maintains an industrial statistics database, which is regularly updated with the data, collected from national statistical offices and OECD (for OECD member countries). UNIDO also collects national accounts-related data from the National Accounts Main Aggregates Database of UNSD, the World Development Indicators of the World Bank and other secondary sources. Such data are primarily used to compile statistics related to Manufacturing Value Added (MVA); its growth rate and share in gross domestic product (GDP) in various countries and regions. UNIDO disseminates industrial data through its publication of the International Yearbook of Industrial Statistics, CD products and through the newly developed online portal at <http://stat.unido.org>.

The statistics team in UNIDO started using R already in 2008, when the migration of the complete statistical system from a mainframe to a client/server architecture was completed. While the main production line is still in SAS[®] and .NET, all new applications and tools are developed in R. UNIDO has published two research papers on the topic *R in the Statistical Office* (Todorov 2010; Todorov and Templ 2012). 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. One example application emphasizes the graphical excellence of R as applied for generating publication quality graphics included in the *International Yearbook of Industrial Statistics* (UNIDO 2014). The graphics together with the related text are typeset in L^AT_EX using the R tool for dynamic reporting **Sweave**. Another application illustrates the analytical and modeling functions available in R and within add-on packages (see Boudt, Todorov, and Upadhyaya 2009). These are used to implement a *nowcasting* tool for Manufacturing Value Added (MVA) to generate estimates for UNIDO publications. Functions from the package for robust statistics **robustbase** (Rousseeuw, Croux, Todorov, Ruckstuhl, Salibian-Barrera, Verbeke, and Maechler 2013) are used for this purpose.

UNIDO has developed several packages for internal use, but some of the packages are also available online. The package **CIttools** provides tools for computation and evaluation of composite indicators. It was developed for computing and analysis of the UNIDO's *Competitiveness Industrial Performance Index* (UNIDO 2013) and is available from *R-Forge* (<https://r-forge.r-project.org/projects/cia/>).

The World Input-Output Database (WIOD) (Timmer, Erumban, Gouma, Los, Temurshoev, de Vries, Arto, Genty, Neuwahl, Rueda-Cantuche, Villanueva, Francois, Pindyuk, Poschl, Stehrer, and Streicher 2012) is a new public data source which provides time-series of world input-output tables for the period from 1995 to 2009. The package **rwiot** developed by UNIDO provides analytical tools for exploration of the various dimensions of the internationalization

of production through time and across countries using input-output analysis. The package contains functions for basic (Leontief and Goshian inverse, backward and forward linkage, impact analysis) as well as advanced (vertical specialization) input-output analysis. Compositional data analysis techniques (Facevicová, Hron, Todorov, Guo, and Templ 2014) can be applied to study the interregional intermediate flows by sector and by region.

UNIDO's technical assistance to developing countries and countries with economies in transition is aimed at either creating a new industrial database or improving the existing statistical system. Through its technical support, UNIDO promotes the quality assurance of industrial statistics by statistical projects that are designed to assist in producing accurate, complete and internationally comparable statistical data. As a main statistical tool for data management and analysis the R environment is promoted. Training materials, example data sets and packages are prepared and courses are carried out, but still a lot of work is necessary to make R the statistical software of choice in the developing countries.

5. Examples

In the following section we provide two examples of R in official statistics. The first one illustrates the estimation of gender pay gap while the second one gives a short overview of anonymization of data.

5.1. Indicators and models from SES

In the European Union the gender pay gap is estimated from the *Structure of Earnings Survey* (SES) which is conducted in nearly all European countries. Similar surveys are collected all over the world. SES data includes information on enterprise and employment level. Generally, such linked employer-employee data are used to identify the determinants/differentials of earnings, but some indicators are also directly derived from hourly earnings, like the gender pay gap or the Gini coefficient (Gini 1912). SES is a complex survey of enterprises and establishments with more than 10 employees (e.g. 11.600 enterprises in Austria), NACE (an European industry standard classification system consisting of a 6 digit code) C-O, including a large sample of employees (in Austria: 207.000). The sampling units can be selected using the **sampling** (Tillé 2006) or the **simFrame** (Alfons *et al.* 2011b) package, for example. In many countries, a two-stage design is used, whereby a stratified sample of enterprises and establishments on the NACE 1-letter section level, *NUTS1* (Nomenclature of Territorial Units for Statistics) and employment size range is used in the first stage, and large enterprises may have higher inclusion probabilities. In stage 2, systematic sampling or simple random sampling is applied on each enterprise to sample employees. Often, unequal inclusion probabilities regarding employment size range categories are included here.

Unit non-responses can be considered by applying calibration through **survey** (Lumley 2010), **sampling** or **ReGenesees** packages. It is also important to understand the behavior of non-response items. To analyze the missing values structure using exploratory and visual tools, the package **VIM** (Templ *et al.* 2012) can be used. **VIM** also provides model-based imputation methods built on robust estimates that can deal with all kinds of variables. For example, one of the functions **irmi** or **kNN** can be used to impute item non-responses in the data. Calibration is applied to represent some population characteristics corresponding to NUTS 2 and NACE 1-digit level, but calibration is also carried out for the gender characteristic (number of males and females in the population). Here, the package **survey**, **sampling** or the **calibWeights** function from the package **laeken** (Alfons and Templ 2013) or **simPop** can be used.

Our example focuses on the gender pay gap as implemented in the R-package **laeken** (Alfons, Holzer, and Templ 2011a; Alfons and Templ 2013). The classical estimates – presented here as breakdown by education – are obtained by the **gpg()** function, assuming that the analyzed data are stored as a data frame. The **print** method for the resulting objects displays the estimates – the overall estimate and the estimates from the chosen breakdown. Not

surprisingly, the sampling weights were specified by the `weights` argument. The code for estimation of the Gender Pay Gap using the package `laeken` is as follows:

```
> library("laeken")
> data("ses")
> g1 <- gpg(inc = "earningsHour", method = "median",
  gender = "sex", weights = "GrossingUpFactor.x",
  breakdown = "education", data = ses)
> g1
g1
Value:
[1] 0.1938192
```

```
Value by stratum:
      stratum      value
1 ISCED 0 and 1 0.2086474
2      ISCED 2 0.1487547
3 ISCED 3 and 4 0.1695580
4      ISCED 5A 0.2974547
5      ISCED 5B 0.2198194
```

The variance of these point estimates are estimated using the syntax below. Here, a calibrated bootstrap is applied for variance estimation ([Alfons and Templ 2013](#)).

```
variance("earningsHour", weights = "GrossingUpFactor.x",
  gender="Sex", data = x, indicator = g1,
  X = calibVars(x$Location), breakdown="education", seed = 123)
```

```
Value:
[1] 0.1938192
```

```
Variance:
[1] 2.078831e-05
```

```
Confidence interval:
      lower      upper
0.2253051 0.2439922
```

```
Value by stratum:
      stratum      value
1 ISCED 0 and 1 0.2086474
2      ISCED 2 0.1487547
3 ISCED 3 and 4 0.1695580
4      ISCED 5A 0.2974547
5      ISCED 5B 0.2198194
```

```
Variance by stratum:
      stratum      var
1 ISCED 0 and 1 1.362429e-03
2      ISCED 2 9.451208e-05
3 ISCED 3 and 4 2.191488e-05
4      ISCED 5A 2.218666e-04
5      ISCED 5B 2.475571e-04
```


Confidence interval by stratum:

	stratum	lower	upper
1	ISCED 0 and 1	0.1772622	0.3200365
2	ISCED 2	0.1525004	0.1888864
3	ISCED 3 and 4	0.1951637	0.2142648
4	ISCED 5A	0.3024255	0.3652755
5	ISCED 5B	0.1728143	0.2363946

However, since the gender pay gap is very sensitive to outliers, it is recommended to apply robust methods (Hulliger, Alfons, Filzmoser, Meraner, Schoch, and Templ 2011). The following code snippet shows how to estimate robustly the gender pay gap using semi-parametric modeling (for details, see, Alfons, Templ, Filzmoser, and Holzer 2011c; Hulliger *et al.* 2011) and replacing the outliers in the tail with estimates from the modeled Pareto distribution.

```
ts <- paretoScale(x$earningsHour, w = x$GrossingUpFactor.x)

# estimate shape parameter
fit <- paretoTail(x$earningsHour, k = ts$k,
  w = x$GrossingUpFactor.x)

# replacement of outliers
earningsHour <- replaceOut(fit)

# fit of the gender pay gap
gpg(earningsHour, weights = x$GrossingUpFactor.x)
```

To illustrate model-based estimations and to show (a very simple) model-fitting functionality of R, we choose a model described in Marsden (2010) which was applied in the PiEP Lissy project (http://cordis.europa.eu/docs/publications/8260/82608181-6_en.pdf). This model is also used in Dybczak and Galuscak (2010). OLS regression models are fitted and the gross log hourly earnings of workers in enterprises are modeled, see below.

```
lm1 <- lm(log(earningsHour) ~ Sex + age + I(age^2) + education +
  Occupation, data=x)
summary(lm1)
```

The predicted values for the hourly earnings can then be used to estimate the gender pay gap. With the `summary()` method, the effect of gender, age, education and occupation on the hourly earnings can be displayed. A bunch of diagnostic plots are available with `plot(lm1)`.

5.2. Anonymization of SES

Typically, statistical data are published on the web in tabular form. However, the tabulated information may allow to identify individuals. To avoid the disclosure of individuals in tabular data, primary and secondary suppressions can be made with the help of the package `sdctable` (Meindl 2014). Once the hierarchies are specified, the function `protectTable()` allows to apply various algorithms to anonymize the tables.

However, researchers from various institutions often need microdata for more detailed analysis. If linkage of the SES data with externally released data sources is successfully based on a number of identifiers (key variables), the intruder will have access to all the information related to a specific corresponding unit in the released data. This means that a subset of critical variables can be exploited to disclose everything about a unit in the data set.

The code listed below shows how to set up the disclosure scenario for the SES data. First an object of class `sdcmicroObj` is created by using function `createSdcObj()` (from package

`sdcMicro`) that includes all information on the disclosure scenario. For example, the disclosure risk of our data corresponding to the key variables is already available in the resulting object. We selected categorical and continuous key variables as well as we specified the variable holding information on the sampling weights.

```
library("sdcMicro")
sdc <- createSdcObj(x,
  keyVars = c('Size', 'age', 'Location', 'economicActivity'),
  numVars = c('earningsHour', 'earnings'),
  weightVar = 'GrossingUpFactor.y')
print(sdc, "freq")
```

Number of observations violating

```
- 2-anonymity: 4979
- 3-anonymity: 11291
```

Percentage of observations violating

```
- 2-anonymity: 2.49 %
- 3-anonymity: 5.65 %
```

```
print(sdc, "risk")
```

```
-----
0 obs. with higher risk than the main part
Expected no. of re-identifications:
4956.19 [ 2.48 %]
-----
```

From this output it is easy to see the large number of unique combinations from cross-tabulating the categorical key variables. However, relative to the size of the data (199909 observation) this number is not that large as it looks at the first view, it is about 2.5% of the observations. Nevertheless, certain number of observations may have a considerable higher individual risk, see the last line in the previous code snippet. For details on risk estimation (also on other methods available in the package) we refer to [Franconi and Polettini \(2004\)](#) and [Templ *et al.* \(2015\)](#). It is necessary to recode some categories of the key variables to receive a lower number of uniqueness as well as to apply local suppression as well as anonymizing the continuous key variables. Functions `globalRecode` and `groupVars` (for global recoding), `localSuppression` (heuristic algorithm performing local suppression), `microaggregation`, and many other methods can be directly applied on the object `sdc`.

As an example, below the code is shown how to add correlated noise ([Brand 2004](#)) to continuously scaled key variables. The parameter `noise` determines how many noise (in percentages) is added.

```
sdc <- addNoise(sdc, noise=100)
```

In the object `sdc` all information about disclosure risk and data utility is saved (see the corresponding print methods from the manual). For further reading we refer to [Templ *et al.* \(2015\)](#) and ([Templ, Kowarik, and Meindl 2014a](#)).

5.3. Accessing international statistical databases with R

When conducting economics studies, like competitiveness analysis or benchmarking, it is necessary to access different sources of data. Many international organizations maintain statistical databases which cover certain types of data: COMTRADE, UNCTAD and WTO for international trade data, World Development Indicators (WDI) from the World bank, World Economic Outlook (WEO) and International Financial statistics (IFS) from the International Monetary Fund (IMF), the Industrial statistics databases (INDSTAT) by UNIDO and many more. Some of these organizations already provide an application programming interface (API) for accessing the data which tremendously facilitates the use of these databases. Here we will consider several examples of accessing such databases using code written in R. For some of these APIs R packages are already available, for others only examples of R code are provided.

World Development Indicators

The flagship publication of the World Bank, *World Development Indicators*, presents a comprehensive collection of cross-country comparable development indicators, compiled from officially-recognized international sources. The database contains more than 1300 time series for more than 200 economies. The countries and areas are presented in more than 30 groups, with data for many indicators going back more than 50 years. The statistical tables are available online and are consistently updated based on revisions to the World Development Indicators database. An API is provided by the World Bank which offers direct access to information contained in the World Bank databases. The data can be accessed by country, by type, by topic and more. The CRAN package **WDI** makes it easy to search and download data from the WDI. The package contains essentially two functions – `WDIsearch()` for searching for data and `WDI()` for downloading the selected data into a data frame. Let us search for example for indicators related to CO2 emissions:

```
library("WDI")
co2list <- WDIsearch("CO2 emissions")
head(co2list)
```

indicator	name
[1,] "EN.ATM.CO2E.CP.KT"	"CO2 emissions from cement production (thousand metric tons)"
[2,] "EN.ATM.CO2E.FF.KT"	"CO2 emissions from fossil-fuels, total (thousand metric tons)"
[3,] "EN.ATM.CO2E.FF.ZS"	"CO2 emissions from fossil-fuels (% of total)"
[4,] "EN.ATM.CO2E.GF.KT"	"CO2 emissions from gaseous fuel consumption (kt) "
[5,] "EN.ATM.CO2E.GF.ZS"	"CO2 emissions from gaseous fuel consumption (% of total) "
[6,] "EN.ATM.CO2E.GL.KT"	"CO2 emissions from gas flaring (thousand metric tons)"

The function `WDIsearch()` uses `grep`, which allows to use (case-insensitive) regular expressions. For example if searching for manufacturing value added at constant prices we could write:

```
WDIsearch("manufacturing.*value.*constant")
```

indicator	name
[1,] "NV.IND.MANF.KD"	"Manufacturing, value added (constant 2005 US\$)"
[2,] "NV.IND.MANF.KN"	"Manufacturing, value added (constant LCU)"

Having identified the indicator we need, e.g. `NV.IND.MANF.KD` for *Manufacturing, value added (constant 2005 US\$)* we can download the data for one or more countries. Let us download also the total population (indicator `'SP.POP.TOTL'`) for the same countries and compute MVA per capita.

```
df.mva <- WDI("NV.IND.MANF.KD", country=c("IN","MY","ID"), start=1970, end=2013)
df.pop <- WDI("SP.POP.TOTL", country=c("IN","MY","ID"), start=1970, end=2013)
df <- merge(df.pop, df.mva)
names(df)[4:5] <- c("POP", "MVA")
df$MVACAP <- round(df$MVA/df$POP)
head(df)
```

	iso2c	country	year	POP	MVA	MVACAP
1	ID	Indonesia	1970	114066887	2750076402	24
2	ID	Indonesia	1971	116996006	2836927088	24
3	ID	Indonesia	1972	119974444	3265345458	27
4	ID	Indonesia	1973	123002081	3763262762	31
5	ID	Indonesia	1974	126080548	4371172606	35
6	ID	Indonesia	1975	129210098	4909605912	38

Figure 4 shows the MVA for the selected countries.

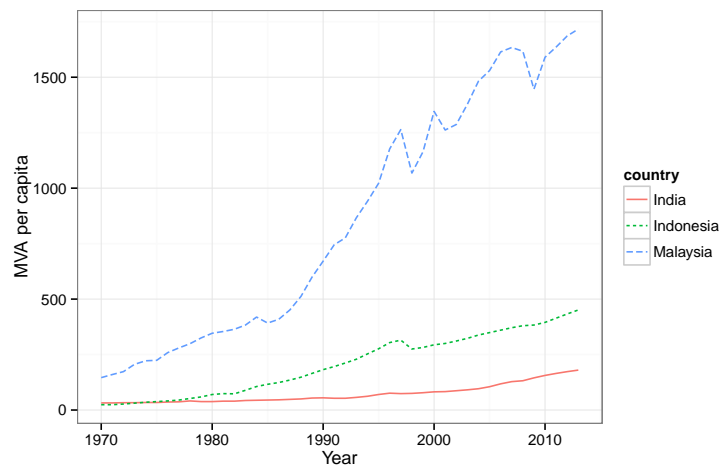


Figure 4: Manufacturing value added (MVA) per capita in India, Indonesia and Malaysia, using data from World Development Indicators.

UN COMTRADE

The United Nations Commodity Trade Statistics Database (UN COMTRADE) contains annual international trade data for over 170 reporter countries, detailed by commodities and partner countries. This is the largest repository of international trade data. All values are converted from national currency into current US dollars using exchange rates supplied either by the reporter countries, or derived from monthly market rates and volume of trade. The data are reported in current classification and revision – Harmonized system (HS) in most cases – and are converted all the way down to the earliest classification SITC revision 1. The time series start as far back as 1962 and go up to the most recent completed year. The data are available from Internet at <http://comtrade.un.org/db/default.aspx>, but recently a new API was developed (still in beta version). A HTTP GET request is sent to the URL <http://comtrade.un.org/api/get?...> and the output may be (currently) in comma-separated values (CSV) or JSON format. In the following will be shown how to use this API for simple data queries. First let us get the list of all reporting countries and areas and find the codes of some countries – which we will use later for retrieving the data.

```
## Read the list of all reporting countries/areas
```

```
library("rjson")
jsonFile <- "http://comtrade.un.org/data/cache/reporterAreas.json"
repCountry <- fromJSON(file=jsonFile)
repCountry <- as.data.frame(do.call(rbind, repCountry[[2]]))
colnames(repCountry) <- c("code", "name")
head(repCountry)
```

```
  code      name
1  all      All
2    4 Afghanistan
3    8      Albania
4   12      Algeria
5   20      Andorra
6   24      Angola
```

```
subset(repCountry, name %in% c("Austria", "Bulgaria"))
```

```
  code      name
13   40 Austria
33  100 Bulgaria
```

The code of Bulgaria is 100 and of Austria is 40. Let us now build a query and retrieve all trade flows (imports and exports) from Bulgaria to Austria in 2010. Only the total for all commodities will be requested. All other parameters remain defaults.

```
comtradeURL <- "http://comtrade.un.org/api/get?"
ps <- "2010"    # one year
r <- 100       # reporting country=Bulgaria
p <- 40        # partner = Austria
rg <- "all"    # trade flow
cc <- "TOTAL"  # total of all commodities

comtradeURL <- paste0(comtradeURL,
                      "ps=", ps, "&",    # time period
                      "r=", r, "&",      # reporting area
                      "p=", p, "&",      # partner country
                      "rg=", rg, "&",    # trade flow
                      "cc=", cc, "&",    # commodities
                      "fmt=csv",        # format is CSV
                      sep = "")

dd <- read.csv(comtradeURL, header=TRUE)
dd
```

```
  Classification Year Period Period.Desc. Aggregate.Level
1              H3 2010   2010         2010             0
2              H3 2010   2010         2010             0
  Is.Leaf.Code Trade.Flow.Code Trade.Flow Reporter.Code Reporter
1           0           1      Import           100 Bulgaria
2           0           2      Export           100 Bulgaria
  Reporter.ISO Partner.Code Partner Partner.ISO Commodity.Code
1           BGR           40 Austria           AUT          TOTAL
2           BGR           40 Austria           AUT          TOTAL
  Commodity Qty.Unit.Code Qty.Unit Qty Netweight..kg.
1 All Commodities           1 No Quantity NA          NA
```

2	All Commodities	1	No Quantity	NA	NA
	Trade.Value..US..	Flag			
1	882147923	0			
2	388469998	0			

Similarly, data can be downloaded from COMTRADE using JSON. Other formats (Microsoft Excel, **SDMX**) are planned. The maximum number of records is limited to 50.000 per data query. For authorized users batch mode download through the API will be available. As already mentioned, this is still an experimental API and for production is recommended to us the legacy API.

6. Summary and conclusions

There is an increasing demand for statistical tools, which combine easy to use traditional software packages with newest analytical methods and one very popular such tool is the statistical programming language R. In this contribution, we briefly described its usefulness in the daily work of statistical offices, listed and briefly presented the most popular R packages for survey methodology.

The development of R packages for specific areas of official statistics is growing quickly. For example, R provides many more methods for statistical disclosure control than any other software package. The situation is similar in indicators methodologies and survey statistics, especially new developments using robust methodology are available (see for example the robust estimation of gender pay gap in the listing in Section 5). The development of such new tools was strongly supported by international activities such as the AMELI project (Münnich, Alfons, Bruch, Filzmoser, Graf, Hulliger, Kolb, Lehtonen, Lussmann, Meraner, Nedyalkova, Schoch, Templ, Valaste, Veijanen, and Zins 2011).

The usefulness and power of various R packages can be shown, whereas the field of application is manifold. First, R can be used to work efficiently with data, either computing in memory using new packages like **dplyr** and **data.table** or by connecting to databases. Specialized packages allow the user-friendly application of R to many specific fields in official statistics and survey methodology, as shown in the example section for survey statistics and remote access to statistical databases. Infrastructure for R is provided in various national and international statistical offices; this includes the distribution of R, the internal organizational support, the development of packages and holding training courses on R. Our contribution gives an outline of the usefulness of R for statisticians, methodologists, subject matter specialists and statistical stakeholders in the area of official statistics and survey methodology.

And finally note that R is not only freeware, free software and open-source, one of the greatest advantages of R is its online support (Tippmann 2015).

Acknowledgment

We would like to thank Ana Maria Dobre for providing us further details on the use of R in Statistics Romania. Special thanks goes to Rainer Stütz for numerous helpful comments and improvements of the text.

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization (UNIDO).

References

Alfons A, Holzer J, Templ M (2011a). **laeken**: *Laeken Indicators for Measuring Social Cohesion*. R package version 0.2.2, URL <http://CRAN.R-project.org/package=laeken>.

- Alfons A, Kraft S, Templ M, Filzmoser P (2011b). “Simulation of Close-to-Reality Population Data for Household Surveys with Application to EU-SILC.” *Statistical Methods & Applications*, **20**(3), 383–407. doi:10.1007/s10260-011-0163-2.
- Alfons A, Templ M (2013). “Estimation of Social Exclusion Indicators from Complex Surveys: The R Package **laeken**.” *Journal of Statistical Software*, **54**(15), 1–25.
- Alfons A, Templ M, Filzmoser P (2010). “An Object-Oriented Framework for Statistical Simulation: The R Package **simFrame**.” *Journal of Statistical Software*, **37**(3), 1–36.
- Alfons A, Templ M, Filzmoser P, Holzer J (2011c). “Robust Pareto Tail Modeling for the Estimation of Indicators on Social Exclusion using the R Package **laeken**.” *Research Report CS-2011-2*, Department of Statistics and Probability Theory, Vienna University of Technology.
- Allaire J, McPherson J, Xie Y, Wickham H, Cheng J, Allen J (2014). **rmarkdown**: *Dynamic Documents for R*. R package version 0.3.3, URL <http://rmarkdown.rstudio.com>.
- Attali D (2016). **shinyjs**: *Perform Common JavaScript Operations in shiny Apps using Plain R Code*. R package version 0.4.0, URL <https://CRAN.R-project.org/package=shinyjs>.
- Bache S, Wickham H (2014). **magrittr**: *A Forward-Pipe Operator for R*. R package version 1.5, URL <http://CRAN.R-project.org/package=magrittr>.
- Baddeley A, Turner R (2005). “**spatstat**: An R Package for Analyzing Spatial Point Patterns.” *Journal of Statistical Software*, **12**(6), 1–42.
- Boonstra H (2012). **hbsae**: *Hierarchical Bayesian Small Area Estimation*. R package version 1.0, URL <http://CRAN.R-project.org/package=hbsae>.
- Borg A, Sariyar M (2015). **RecordLinkage**: *Record Linkage in R*. R package version 0.4-7, URL <http://CRAN.R-project.org/package=RecordLinkage>.
- Boudt K, Todorov V, Upadhyaya S (2009). “Nowcasting Manufacturing Value Added for Cross-Country Comparison.” *Statistical Journal of the IAOS: Journal of the International Association of Official Statistics*, **26**, 15–20.
- Brand R (2004). “Microdata Protection Through Noise Addition.” In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, pp. 347–359. Springer, Berlin, Heidelberg.
- Breidenbach J (2013). **JoSAE**: *Functions for Some Unit-Level Small Area Estimators and their Variances*. R package version 0.2.2, URL <http://CRAN.R-project.org/package=JoSAE>.
- Brown G, Mills J, Ayoubkhani D, Gallop A (2013). “Smoothing Mortality Rates Using R.” *Research report*, ONS.
- Camarda C (2012). “**MortalitySmooth**: An R Package for Smoothing Poisson Counts with P-Splines.” *Journal of Statistical Software*, **50**(1), 1–24. ISSN 1548-7660.
- de Jonge E, van der Loo M (2012). **editrules**: *R Package for Parsing and Manipulating Edit Rules*. R package version 2.2-0, URL <http://CRAN.R-project.org/package=editrules>.
- Dinges G, Kowarik A, Meindl B, Templ M (2011). “An Open Source Approach for Modern Teaching Methods: The Interactive TGUI System.” *Journal of Statistical Software*, **39**(7), 1–19.
- Dobre A, Adam R (2014). “The Progress of R in Romanian Official Statistics.” *Romanian Statistical Review*, **2**, 45–54.

- D’Orazio M, Di Zio M, Scanu M (2006). *Statistical Matching: Theory and Practice*. Wiley Series in Survey Methodology. John Wiley & Sons, Chichester, England; Hoboken, NJ. ISBN 9780470023549.
- Dowle M, Short T, Lianoglou S, Srinivasan A (2014). **data.table**: *Extension of data.frame*. R package version 1.9.2, URL <http://CRAN.R-project.org/package=data.table>.
- Dybczak K, Galuscak K (2010). “Changes in the Czech Wage Structure: Does Immigration Matter?” *Working paper series no 1242*, European Central Bank. Wage dynamics network.
- Facevicová K, Hron K, Todorov V, Guo D, Templ M (2014). “Logratio Approach to Statistical Analysis of 2x2 Compositional Tables.” *Applied Statistics*, **41**(5), 944–958.
- Fox J (2005). “The R Commander: A Basic Statistics Graphical User Interface to R.” *Journal of Statistical Software*, **14**(9), 1–42.
- Franconi L, Poletti S (2004). “Individual Risk Estimation in μ -Argus: A Review.” In J Domingo-Ferrer (ed.), *Privacy in Statistical Databases, Lecture Notes in Computer Science*, pp. 262–272. Springer, Berlin, Heidelberg.
- Gentlemen R (2009). “Data Analysts Captivated by R’s Power.” URL <http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html>.
- Gini C (1912). “Variabilità E Mutabilità: Contributo Allo Studio Delle Distribuzioni E Delle Relazioni Statistiche.” *Studi Economico-Giuridici della R. Università di Cagliari*, **3**, 3–159.
- Godfrey A (2013). “Statistical Analysis from a Blind Person’s Perspective.” *The R Journal*, **5**(1), 73–80.
- Honaker J, King G, Blackwell M (2011). “Amelia II: A Program for Missing Data.” *Journal of Statistical Software*, **45**(7), 1–47.
- Horner J (2011). **brew**: *Templating Framework for Report Generation*. R package version 1.0-6, URL <http://CRAN.R-project.org/package=brew>.
- Hulliger B, Alfons A, Filzmoser P, Meraner A, Schoch T, Templ M (2011). “Robust Methodology for Laeken Indicators.” *Research Project Report WP4 – D4.2*, FP7-SSH-2007-217322 AMELI. URL <http://ameli.surveystatistics.net>.
- Kowarik A, Meindl B, Templ M (2014a). **sparkTable**: *Sparklines and Graphical Tables for T_EX and HTML*. R package version 0.12.0, URL <http://CRAN.R-project.org/package=sparkTable>.
- Kowarik A, Meraner A, Templ M, Schopfhauser D (2014b). “Seasonal Adjustment with the R Packages **x12** and **x12GUI**.” *Journal of Statistical Software*, **62**(2), 1–21.
- Kowarik A, Templ M, Meindl B, Fontenau F (2014c). “Graphical User Interface for Package **sdcmicro**.” *Technical report*, International Household Survey Network. URL <http://www.ihsn.org/home/sites/default/files/resources/sdcMicroGUI.pdf>.
- Leisch F (2003). “**Sweave**, Part II: Package Vignettes.” *R News*, **3**(2), 21–24.
- Leisch F, Rossini A (2003). “Reproducible Statistical Research.” *Chance*, **16**(2), 46–50.
- Lumley T (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley, Hoboken, NJ. ISBN 9780470284308.
- Marsden D (2010). “Pay Inequalities and Economic Performance.” *Technical Report PiEP Final Report V4*, Centre for Economic Performance London School of Economics, London.

- Meindl B (2014). **sdcTable**: *Methods for Statistical Disclosure Control in Tabular Data*. R package version 0.13.0, URL <http://CRAN.R-project.org/package=sdcTable>.
- Meindl B, Templ M, Alfons A, Kowarik A (2014). **simPop**: *Simulation of Synthetic Populations for Surveys Based On Auxiliary Data*. R package version 0.2.6, URL <http://CRAN.R-project.org/package=simPop>.
- Mirai Solutions GmbH (2015). **XLConnect**: *Excel Connector for R*. R package version 0.2-11, URL <http://CRAN.R-project.org/package=XLConnect>.
- Münnich R, Alfons A, Bruch C, Filzmoser P, Graf M, Hulliger B, Kolb JP, Lehtonen R, Lussmann D, Meraner A, Nedyalkova D, Schoch T, Templ M, Valaste M, Veijanen A, Zins S (2011). “Policy Recommendations and Methodological Report.” *Research Project Report WP10 – D10.1/D10.2*, FP7-SSH-2007-217322 AMELI. URL <http://ameli.surveystatistics.net>.
- Petris G (2010). “An R Package for Dynamic Linear Models.” *Journal of Statistical Software*, **36**(12), 1–16.
- Radinger R, Nachtmann G, Peterbauer J, Reif M, Hanika A, Kowarik A, Lehner D (2014). *Hochschulprognose 2014*. URL http://www.statistik.at/web_de/static/hochschulprognose_2014_063538.pdf.
- R Core Team (2015). **foreign**: *Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, . . .*. R package version 0.8-63, URL <http://CRAN.R-project.org/package=foreign>.
- Rousseeuw PJ, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Maechler M (2013). **robustbase**: *Basic Robust Statistics*. R package version 0.4-5, URL <http://CRAN.R-project.org/package=robustbase>.
- RStudio Inc (2014). **shiny**: *Web Application Framework for R*. R package version 0.10.2.1, URL <http://CRAN.R-project.org/package=shiny>.
- Sarkar D (2008). **lattice**: *Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.
- Schafer J (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
- Schoch T (2014). “Robust Unit-Level Small Area Estimation: A Fast Algorithm for Large Datasets.” *Austrian Journal of Statistics*, **41**(4), 243–265.
- Schopfhaue D, Templ M, Alfons A, Kowarik A, Prantner B (2014). **VIMGUI**: *Visualization and Imputation of Missing Values*. R package version 0.9.0, URL <http://CRAN.R-project.org/package=VIMGUI>.
- Templ M, Alfons A, Filzmoser P (2012). “Exploring Incomplete Data Using Visualization Techniques.” *Advances in Data Analysis and Classification*, **6**(1), 29–47. doi:10.1007/s11634-011-0102-y.
- Templ M, Hulliger B, Kowarik A, Fürst K (2013). “Combining Geographical Information and Traditional Plots: The Checkerplot.” *International Journal of Geographical Information Science*, **27**(4), 685–698.
- Templ M, Kowarik A, Filzmoser P (2011). “Iterative Stepwise Regression Imputation using Standard and Robust Methods.” *Computational Statistics & Data Analysis*, **55**(10), 2793–2806.

- Templ M, Kowarik A, Meindl B (2014a). “Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation.” *Technical report*, International Household Survey Network. URL <http://www.ihsn.org/home/sites/default/files/resources/sdcMicro.pdf>.
- Templ M, Kowarik A, Meindl B (2015). “Statistical Disclosure Control for Micro-Data Using the R Package **sdcMicro**.” *Journal of Statistical Software*, **67**(4), 1–36. doi:10.18637/jss.v067.i04.
- Templ M, Meindl B (2010). “Practical Applications in Statistical Disclosure Control Using R.” In J Nin, J Herranz (eds.), *Privacy and Anonymity in Information Management Systems*, Advanced Information and Knowledge Processing, pp. 31–62. Springer, London.
- Templ M, Meindl B, Kowarik A (2014b). “Tutorial for **sdcMicroGUI**.” *Technical report*, International Household Survey Network. URL <http://www.ihsn.org/home/sites/default/files/resources/Tutorial%20sdcMicroGUI%20v6.pdf>.
- Templ M, Meindl B, Kowarik A, Chen S (2014c). “Introduction to Statistical Disclosure Control (SDC).” *Technical Report IHSN Working Paper No 007*, International Household Survey Network. URL <http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf>.
- Temple Lang D (2013). *XML: Tools for Parsing and Generating XML within R and S-Plus*. R package version 3.98-1.1, URL <http://CRAN.R-project.org/package=XML>.
- Tennekes M, de Jonge E, Daas P (2013). “Visualizing and Inspecting Large Datasets with Tableplots.” *Journal of Data Science*, **11**(1), 43–58.
- Tillé Y (2006). *Sampling Algorithms*. Springer Series in Statistics. Springer, New York. ISBN 9780387308142.
- Timmer M, Erumban AA, Gouma R, Los B, Temurshoev U, de Vries GJ, Arto I, Genty VAA, Neuwahl F, Rueda-Cantuche JM, Villanueva A, Francois J, Pindyuk O, Poschl J, Stehrer R, Streicher G (2012). “The World Input-Output Database (WIOD): Contents, Sources and Methods.” WIOD Background document, URL www.wiod.org.
- Tippmann S (2015). “Programming tools: Adventures with R.” *Nature*, pp. 109–110. doi:10.1038/517109a.
- Todorov V (2010). “R in the Statistical Office: The UNIDO Experience.” *Working Paper 03/2010 1*, United Nations Industrial Development.
- Todorov V, Templ M (2012). “R in the Statistical Office: Part II.” *Working paper 1/2012*, United Nations Industrial Development.
- Todorov V, Templ M, Filzmoser P (2011). “Detection of Multivariate Outliers in Business Survey Data with Incomplete Information.” *Advances in Data Analysis and Classification*, **5**(1), 37–56.
- Tufte ER (2001). *The Visual Display of Quantitative Information*. 2nd edition. Graphics Press, Cheshire, CT. ISBN 0961392142.
- UNIDO (2013). “The Industrial Competitiveness of Nations: Competitive Industrial Performance Report 2012/2013.” *Technical report*, UNIDO, Vienna.
- UNIDO (2014). *International Yearbook of Industrial Statistics*. Edward Elgar Publishing Ltd, Glensanda House, Montpellier Parade, Cheltenham Glos GL50 1UA, UK.

- van Buuren S, Groothuis-Oudshoorn K (2011). “**mice**: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software*, **45**(3), 1–67. URL <http://www.jstatsoft.org/v45/i03/>.
- van der Loo M (2012). “The Introduction and use of R Software at Statistics Netherlands.” In *Proceedings of the Third International Conference of Establishment Surveys (CD-ROM)*. American Statistical Association, Montréal, Canada. URL <http://www.amstat.org/meetings/ices/2012/papers/302187.pdf>.
- van der Loo M (2014). “The **stringdist** Package for Approximate String Matching.” *The R Journal*, **6**, 111–122.
- Warnholz W, Schmid T (2015). **saeSim**: *Simulation Tools for Small Area Estimation*. R package version 0.7.0, URL <http://CRAN.R-project.org/package=saeSim>.
- Wickham H (2009). **ggplot2**: *Elegant Graphics for Data Analysis*. Springer, New York. ISBN 978-0-387-98140-6.
- Wickham H (2015a). **readxl**: *Read Excel Files*. R package version 0.1.0, URL <http://CRAN.R-project.org/package=readxl>.
- Wickham H (2015b). **xml2**: *Parse XML*. R package version 0.1.0, URL <http://CRAN.R-project.org/package=xml2>.
- Wickham H, Francois F (2014). **dplyr**: *A Grammar of Data Manipulation*. R package version 0.2.0.9000, URL <https://github.com/hadley/dplyr>.
- Wickham H, Francois R (2015). **readr**: *Read Tabular Data*. R package version 0.1.0, URL <http://CRAN.R-project.org/package=readr>.
- Wickham H, Miller E (2015). **haven**: *Import SPSS, Stata and SAS Files*. R package version 0.2.0, URL <http://CRAN.R-project.org/package=haven>.
- Wilkinson L, Wills G (2005). *The Grammar of Graphics*. Springer, New York. ISBN 0387245448.
- Xie Y (2013). *Dynamic Documents with R and knitr*. Chapman & Hall/CRC The R Series. Taylor & Francis. ISBN 9781482203530.
- Yu-Sung S, Gelman A, Hill J, Yajima M (2011). “Multiple Imputation with Diagnostics (Mi) in R: Opening Windows into the Black Box.” *Journal of Statistical Software*, **45**(2), 1–31.
- Zeileis A (2014). **ineq**: *Measuring Inequality, Concentration, and Poverty*. R package version 0.2-13, URL <http://CRAN.R-project.org/package=ineq>.

Affiliation:

Matthias Templ
CSTAT – Computational Statistics
Institute of Statistics & Mathematical Methods in Economics
Vienna University of Technology
Wiedner Hauptstr. 8–10
1040 Vienna, Austria
Tel. +43 1 58801 10562
e-mail: matthias.templ@tuwien.ac.at
<http://institute.tuwien.ac.at/cstat/>

Valentin Todorov
United Nations Industrial Development Organization (UNIDO),
Vienna International Centre, P.O. Box 300,
A-1400 Vienna, Austria
e-mail: v.todorov@unido.org