Development and Current Practice in Using R at Statistics Austria

Matthias TEMPL
Statistics Austria, Vienna University of Technology
Alexander KOWARIK
Bernhard MEINDL
Statistics Austria

Abstract: The popularity of R is increasing in national statistical offices not only for simulation tasks. Nowadays R is also used in the production process. A lot of new features for various tasks in official statistics have been developed over the last years and these features are freely available in the form of add-on package.

In this contribution we first give an outline of the use of R at Statistics Austria. Discussed is the necessary infrastructure according to the R-installation, the teaching of employees and the support provided to the staff who use R in their daily work.

In the second part, the R developments from the methods unit at Statistics Austria are summarised. The developed packages include methods for data pre-processing (e.g imputation) up to packages for the final dissemination of data including packages for statistical disclosure control, estimation of indicators and the visualisation of results.

Keywords: Official Statistics, Computational Statistics, R

JEL Classification: C630 Computational Techniques; Simulation Modeling [1] [2]

1 R Software Features

R is a free and open-souce environment for statistical computing and graphics. It includes a well-structured function- and object-oriented programming language. Nowadays, R is already the state-of-the-art software for statistical computing in academics but also gains importance in statistical offices as well as in private enterprises.

R is termed an environment because it features beside well-developed functionality for data manipulation, operators for calculations with vectors, matrices, arrays and tools for data analysis and graphics. Additionally, the interaction with other well established software packages is a major strength of R:

- interfaces to other programming languages such as C, C++, Java or Python
- excellent import/export tools for data exchange in csv, excel, SDMX, XML, Stata, SPSS, SAS (Xport sas7bdat), JSON, fixed width format, binary formats
- functions that allow connections to important databases, e.g. DB2 (ODBC, JDBC), MySQL, PostgreSQL, Oracle

In R functions, classes and methods can be defined and created by users, this provides much more freedom and flexibility than for example a macro languages. It should be mentioned that users have access to the same tools as developers. This is one of the reasons why there are almost 6000 add-on-packages ready to be downloaded from the comprehensive R archive network (CRAN).

2 Policy of Using R in Statistics Austria

R is an open source project and the support is - just as with the tradition of other open source projects - given by the community. Since the R-community is large and very strong, chances are high that the community detects possible bugs in packages in due time, and - in contradiction to most commercial software developers - bugs are fixed soon.

Without doubt R is one of the most used software in academics to teach students in statistics. Thus, many former students that start working in official statistics or in companies are familiar with R and want to use it when they start to work in statistical agencies, see also van der Loo [2012], Gentlemen [2009].

2.1 Infrastructure at Statistics Austria

R is available for almost all operating systems including the current and last versions of Windows, OSX and all popular Linux distributions. At Statistics Austria, R is currently installed on more than 60 computers (Windows 7 platform) and on powerful virtual servers featuring a *PowerPPC* architecture and *SUSE Linux Enterprise Server*. The server solution is mainly used for tasks that involve large memory requirements or use multiple cores in parallel.

The leading R-Team at Statistics Austria consists of three experts from the methods division. In addition, each department has chosen one person as first contact person for questions and problems that can easily be answered.

Furthermore, the following organizational setup is in place:

- the R-experts at the methods unit (the administrators) take care of the version of R, they decide on the GUI-front-end and the packages installed with the default installation. All necessary information and files (R, Rstudio[1], packages, documentation and examples) are placed on a particular server.

- the IT department takes these files and deploys the R-installation including the front-end (RStudio) to users. This ensures that only one standardised software package is installed on all computers;
- the general R-support is centralised 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 information in a knowledge infrastructure;
- the administrators define access rights for users on the servers, the mailing list, wiki, file depot, etc. At the moment, basically two user groups exist:
  1. R administrators:
     they have read and write access and are responsible for the folder that contains the entire 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 documentation and read and write access to the wiki and are members of a R-User mailing list.

3 Education

Tbere are currently two courses about R offered to employees at Statistics Austria.

3.1 Basic Course

The basic course is scheduled for 15 hours (5 x 3h) with the aim to bring all participants to a certain level of knowledge. The target group not only consists of beginners, but also of regular R users who learned R in self-study. For the later group, many fundamental insights to the software are presented which are mostly new even for experienced users.

The course consists of the topics data types, import/export (including data base connections), syntax, data manipulation (including the presentation of important add-on packages such as *plyr* and *data.table*) and basic object-orientation features. Ex-cathedra teaching is followed by exercises for the students and R sessions in which the trainers interactively give additional insights.

3.2 Advanced Course

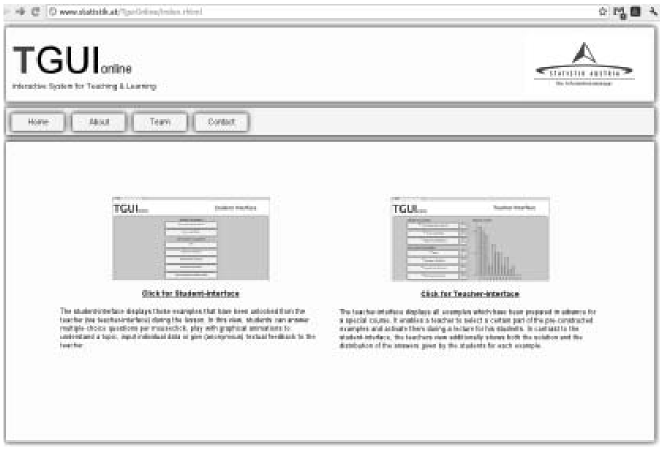
The advanced training also consists of 15 hours with the aim to teach some more complex topics about R.

[1] http://www.rstudio.org

The course consists of the topics graphics (*graphics, grid, lattice, ggplot2*), classes and object-orientation (S3 plus brief introduction to S4 classes), dynamic reporting (*Sweave, knitr, brew, markup*), R development issues (profiling, debugging, benchmarking, basic packaging), web-applications (*shiny*) and gives an overview of other useful packages for key-tasks in official statistics. Again, ex-cathedra teaching is followed by exercises and interactive sessions.

3.3 Usage of R in Methodological Courses

At Statistics Austria, methodological training is offered to the staff. In these courses, R is intensively used for teaching purposes in a way that participants do not get in direct contact with R. The reason is that for these courses no requirements with respect to any software or programming skills should be necessary. Thus, a blended learning system was developed. At the begin of a course, participants fill out an online questionnaire. The collected data is then automatically used in the exercises and is also incorporated into the presentation slides. Participants are able to identify their data in various graphics, tables and other output. Approx. after 20 minutes lecturing, participants have to do exercises using point and click directly in the browser. An online server-client 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 and anonymously saved on the server and aggregated statistics (feedback) are generated automatically with the aim that both the trainers and the participants get an overview if the examples were correctly solved.


www.statistik.at/TguiOnline/index.html
TGUIonline
Interactive System for Teaching & Learning
STATISTIK AUSTRIA
Home
About
Team
Contact
TGUI
Student-Interface
TGUI
Teacher-Interface
Click for Student-interface
Click for Teacher-interface

The student-interface displays those examples that have been unlocked from the teacher (via teacher-interface) during the lesson. In this view, students can answer multiple-choice questions per mouseclick, play with graphical animations to understand a topic, input individual data or give (anonymous) textual feedback to the teacher.

The teacher-interface displays all examples which have been prepared in advance for a special course. It enables a teacher to select a certain part of the pre-constructed examples and activate them during a lecture for his students. In contrast to the student-interface, the teachers view additionally shows both the solution and the distribution of the answers given by the students for each example.

Figure 3.1 Main view of the blended learning system *TGUIonline*.

Figure 3.1 shows the start screen of the developed teaching system in which two different views are implemented. In the teacher-interface, trainers can activate certain examples which are then available to the course participants. They also get feedback how many participants have already solved the question and the correctness of the solutions. In the student-interface, the currently available exercises are listed and can be started.

4 Packages for Official Statistics

4.1 R Task View on Official Statistics

The CRAN Task View *on Official Statistics and Survey Methodology* lists and briefly describes relevant packages that can be used for important tasks in official statistics. The following topics are considered:

- complex survey design;
- editing and visual inspection of microdata;
- imputation;
- statistical disclosure control;
- seasonal adjustment;
- statistical record matching;
- small area estimation;
- indices and indicators.

We refer to the CRAN Task view for further information:
http://cran.r-project.org/web/views/OfficialStatistics.html

4.2 Packages Developed by the Methods Unit at Statistics Austria

The methods unit at Statistics Austria has been using R since 2004. Until recently, SAS was the only software allowed in the statistical production process, but nowadays R starts to replace SAS in tasks, especially with need of modern and emerging methods.

The aim is to implement new methods in R and to provide the developed packages to various projects with the goal to increase the usage of R and decrease the dependence on SAS.

keep on using R for various projects and write code for new projects in R only.

The following packages have been developed by the methods unit (partially together with other organisations):

- `sdcMicro` [Templ et al., 2014], `sdcMicroGUI` [Kowarik et al., 2013b], `sdcTable` [Meindl, 2013]: packages for statistical disclosure control;
- `TGUI`: blended-learning software for teaching;
- `VIM` [Templ et al., 2012]: visualisation and imputation of missing values;
- `x12` [Kowarik and Meraner, 2014a], `x12GUI` [Kowarik and Meraner, 2014b]: batch processing and interactive visualisation of X12-ARIMA
- `sparkTable` [Kowarik et al., 2013a]: sparklines in R for tables with graphics for LaTeX and websites

```
> gini("eqIncome", weights = "rb050", breakdown="db040",
>        data = eusilc)

Value:
[1] 26.48962

Value by stratum:
         stratum    value
1     Burgenland 32.05489
2      Carinthia 25.49448
3  Lower Austria 25.93737
4       Salzburg 25.01652
5         Styria 23.71190
6          Tyrol 25.24881
7  Upper Austria 25.49202
8         Vienna 28.94944
9     Vorarlberg 28.74120
```

Listing 1 Example from package laeken. Estimation of the gini coeficient with breakdown on regions. More advanced features like robust estimation and variance estimation are included in the package but not shown here.

- `laeken` [Alfons and Templ, 2013]: point and variance estimation of poverty indicators;
- `robCompositions` [Templ et al., 2011a]: statistical methods for compositional data.

A short overview about selected packages is now given.

4.2.1 R Package laeken

Units sampled from finite populations typically feature unequal sampling weights, this has to be taken into account when indicators have to be estimated. Additionally, many indicators are non-robust and suffer from strong influence of outliers, which are present in virtually all real-world data sets. The `R` package `laeken` [Alfons and Templ, 2013] is an object-oriented methodological and computational framework for the estimation of indicators from complex survey samples via standard or robust methods. It provides a class structure to allow for easy handling of the functions and objects. Some widely used social exclusion and poverty indicators are implemented together with a calibrated bootstrap framework to estimate the variance of indicators for common survey designs. An example is shown in Listing 1 in which the Gini coefficient is estimated for some regions. The application of more advanced methods (robustification, variance-estimation and plots) is shown in Alfons and Templ [2013], Alfons et al. [2013].

4.2.2 R Package sparkTable

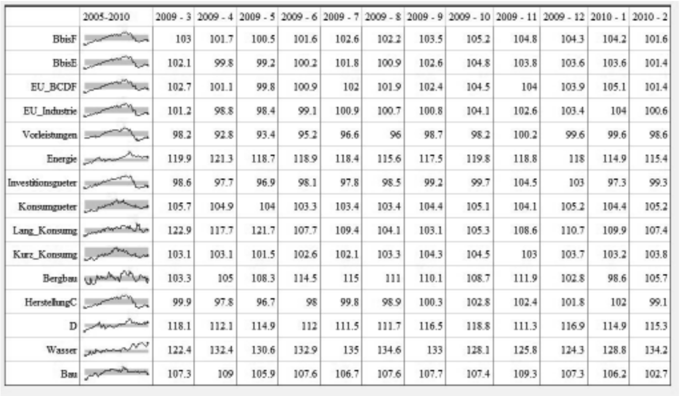
Package `sparkTable` [Kowarik et al., 2013a] provides additional insights into text and tables by the use of small graphics (sparks) in text and graphical tables. Using `sparkTable`, sparklines (time series) , boxplots and bar charts can be produced. Finetuning is possible

```
> sdc <- primarySuppression(sdc, type = "freq", maxN = 10)
> resHYPER <- protectTable(sdc, method="HYPERCUBE")
```

Listing 2 Example code from sdcTable. Primary and secondary cell suppression of an object of class sdcProblem using the hypercube method.

by highlighting specific values, changing colours and or by including statistics in the graphics (e.g. interquantile range), e.g.
.

Figure 4.1 shows the use of sparklines in a graphical table.



	2005-2010	2009 - 3	2009 - 4	2009 - 5	2009 - 6	2009 - 7	2009 - 8	2009 - 9	2009 - 10	2009 - 11	2009 - 12	2010 - 1	2010 - 2
BbisF		103	101.7	100.5	101.6	102.6	102.2	103.5	105.2	104.8	104.3	104.2	101.6
BbisE		102.1	99.8	99.2	100.2	101.8	100.9	102.6	104.8	103.8	103.6	103.6	101.4
EU_BCDF		102.7	101.1	99.8	100.9	102	101.9	102.4	104.5	104	103.9	105.1	101.4
EU_Industrie		101.2	98.8	98.4	99.1	100.9	100.7	100.8	104.1	102.6	103.4	104	100.6
Vorleistungen		98.2	92.8	93.4	95.2	96.6	96	98.7	98.2	100.2	99.6	99.6	98.6
Energie		119.9	121.3	118.7	118.9	118.4	115.6	117.5	119.8	118.8	118	114.9	115.4
Investitionsgueter		98.6	97.7	96.9	98.1	97.8	98.5	99.2	99.7	104.5	103	97.3	99.3
Konsumgueter		105.7	104.9	104	103.3	103.4	103.4	104.4	105.1	104.1	105.2	104.4	105.2
Lang_Konsumg		122.9	117.7	121.7	107.7	109.4	104.1	103.1	105.3	108.6	110.7	109.9	107.4
Kurz_Konsumg		103.1	103.1	101.5	102.6	102.1	103.3	104.3	104.5	103	103.7	103.2	103.8
Bergbau		103.3	105	108.3	114.5	115	111	110.1	108.7	111.9	102.8	98.6	105.7
HerstellungC		99.9	97.8	96.7	98	99.8	98.9	100.3	102.8	102.4	101.8	102	99.1
D		118.1	112.1	114.9	112	111.5	111.7	116.5	118.8	111.3	116.9	114.9	115.3
Wasser		122.4	132.4	130.6	132.9	135	134.6	133	128.1	125.8	124.3	128.8	134.2
Bau		107.3	109	105.9	107.6	106.7	107.6	107.7	107.4	109.3	107.3	106.2	102.7

Figure 4.1 Graphical table produced by sparkTable showing monthly production indices from 2005 till 2010.

4.2.3 R Package sdcTable

The `sdcTable` package [Meindl, 2013] provides methods to generate instances of multidimensional, hierarchical table structures, identify primary sensitive table cells within such objects and finally protect primary sensitive table cells by solving the secondary cell suppression problem with currently three implemented algorithms. First an object of class `sdcProblem` needs to be created. In this step, all possible hierarchies have to be defined and specifics of the table must be listed. After the creation of such an object, the application of primary and secondary cell suppression methods is straightforward to use (see Listing 2, in which primary and secondary cell suppression is applied to object 'sdc' that contains the table that was specified by the user).

4.2.4 R Packages sdcMicro + sdcMicroGUI

The `R` package sdcMicro [Templ et al., 2014] serves as an easy-to-handle, object-oriented S4 class implementation of SDC methods to evaluate and anonymize confidential micro-data sets. All popular disclosure risk and perturbation methods are included. Furthermore, frequency counts, individual

```
> require("sdcMicro"); data("testdata")
> sdc <- createSdcObj(testdata,
>   keyVars=c('urbrur','water','sex','age'),
>   numVars=c('expend','income','savings'),
>   pramVars=c("walls"), w='sampling_weight', hhId='ori_hid')
> print(sdc, "risk")

--------------------------
0 obs. with higher risk than the main part
Expected no. of re-identifications:
 24.78 [ 0.54 %]
--------------------------
--------------------------
Hierarchical risk
--------------------------
Expected no. of re-identifications:
 117.2 [ 2.56 %]

> sdc <- localSuppression(sdc)
> sdc <- microaggregation(sdc)
```

Listing 3 Example from package sdcMicro. Creating an object of class sdcMicroObj and applying local suppression to achieve k-anonymity and microaggregation.

and global risk measures, information loss and data utility statistics are updated (re-calculated) after each anonymization step automatically.

All methods are highly optimized in terms of computational speed. It is possible to work with large data sets like large survey data from India. Reporting facilities that summarize the anonymization process can also easily be used by subject matter specialists and also helps to be reproducible.

In Listing 3, the package is shown in action. An object of class `sdcMicroObj` is created first, then the risk is printed and local suppression and microaggregation are applied to this object. Automatically, the risk and utility measures are updated and all this information is saved in the object. The `sdcMicroGUI` package [Kowarik et al., 2013b] is especially useful for users with limited knowledge in `R` but also `R` experts may use its recoding of variables facilities.

4.2.5 R Packages VIM and VIMGUI

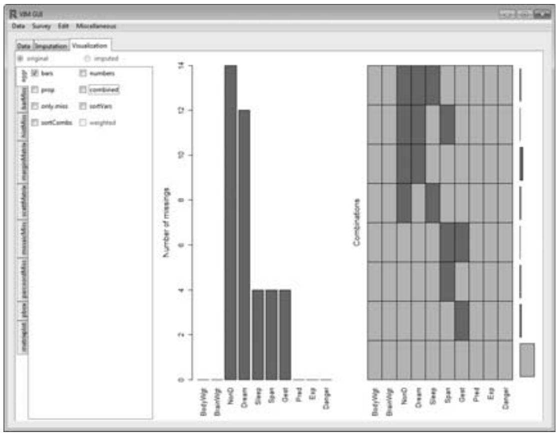
The package `VIM` [Templ et al., 2012] contains visualization techniques to explore the structure of non-complete data sets. Thus, it is not just possible to analyse the structure and relations of missing and non-missing data parts, but also to analyse imputed data. The visualization techniques for missing values are described in detail in [Templ et al., 2012].

In addition, in `VIM` various kinds of imputation methods are included. The choice of available methods is quite extensive and ranges from old-fashioned methods like hot-deck imputation to quite sophisticated methods such as iterative step-wise robust regression imputation [Templ et al., 2011b]. The package can also deal with survey objects from `R` package `survey`.

The package `VIMGUI` is a point and click graphical user interface based on `VIM`. A screenshot of a simple plot is shown in Figure 4.2.

```
> data(testdata); x <- testdata$wna
> imp <- irmi(x, mixed=c("m1", "m2")
```

Listing 4 Example from package VIM applying model-based iterative robust imputation on data including missing values. Semi-continuous variables have to be specified (parameter mixed).



● original ○ imputed

☑ bars ☐ numbers
☐ prop ☐ combined
☐ only.miss ☐ sortVars
☐ sortCombs ☐ weighted

Figure 4.2 Screenshot from the VIMGUI package. Simple aggregation statistics. In the left plot, the number of missing values for each variable are shown while on the right the pattern structure of missing values are displayed.

```
> s <- new("x12Single", ts=AirPassengers)
> s <- setP(s, list(arima.model=c(2,1,1), arima.smodel=c(2,1,1)))
> result <- x12(s)
```

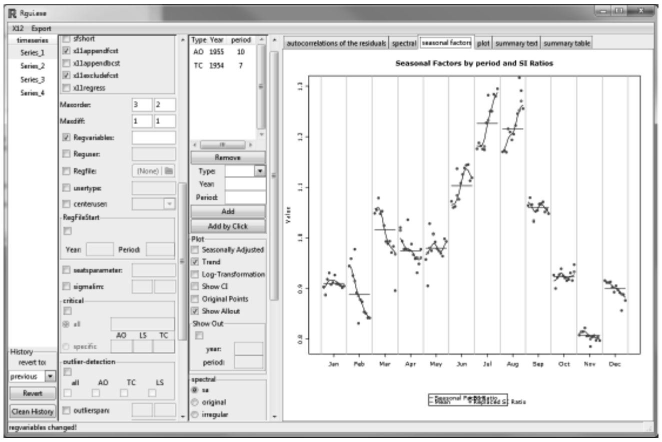
Listing 5 Simple example from package x12. An x12 object is created plus some parameters are set and finally x12 is called with these parameter settings.

4.2.6 R Packages x12 and x12GUI

Different components (mainly: seasonal component, trend component, outlier component and irregular component) of a monthly or quarterly time series can be extracted and a moving holiday effect, a trading day effect and user-defined regressors can be estimated. Computational basis is the `X-12-ARIMA` seasonal adjustment software of the U.S. Census Bureau. The `x12` package [Kowarik and Meraner, 2014a] calls and extracts the output from `X-12-ARIMA` and prepares the resulting output for further processing. The packages serves as an abstraction layer for batch processing `X-12-ARIMA`. New facilities for marking outliers, batch processing and change tracking make the package a powerful and functional tool.

In Listings 5 a simplified example is shown. A single time series is chosen, parameters are specified and the object is finally evaluated. For the resulting object, print, summary and various plot methods are available.

With the `x12GUI` package [Kowarik and Meraner, 2014b] users can interactively select additive outliers, level shifts and temporary changes and the impact is visible immediately. Figure 4.3 shows one view of `x12GUI`.



Type	Year	period
AO	1955	10
TC	1954	7

- □ sfshort
- ☑ x11appendfcst
- □ x11appendbcst
- ☑ x11excludefcst
- □ x11regress

Maxorder: 3 2

Maxdiff: 1 1

- ☑ Regvariables:
- □ Reguser:
- □ Regfile: (None)
- □ usertype:
- □ centeruser:

Plot

- □ Seasonally Adjusted
- ☑ Trend
- □ Log-Transformation
- □ Show CI
- □ Original Points
- ☑ Show Allout

spectral

- ● sa
- ○ original
- ○ irregular

History revert to: previous

regvariables changed!

Figure 4.3 View of one window of the x12GUI package for seasonal adjustment.

5 Conclusion

Nowadays, most new employees with academical background in statistics already have skills in R and are highly motivated to continue using R. By creating an infrastructure for R including training and support, the usage of R in the statistical production process is feasible and often preferable to other software solutions.

The use of specific R-packages related to methods from official statistics makes it possible to tackle problems that are not (easily) solvable with other statistical software packages. This includes survey sampling, calibration, editing, imputation, disclosure control as well as estimation and visualisation.

For some of these packages we provided brief information and we showed simplified examples. The aim was that interested readers become aware of the power of R in official statistics.

New collaborations between countries seem possible since everybody can use the packages for free. Intellectual rights, however, should be respected.

References

A. Alfons and M. Templ. Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, 54(15):1–25, 9 2013.

A. Alfons, M. Templ, and P. Filzmoser. Robust estimation of economic indicators from survey samples based on pareto tail modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(2):271–286, 2013. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2012.01063.x. URL http://dx.doi.org/10.1111/j.1467-9876.2012.01063.x.

R. Gentlemen. Data analysts captivated by R' power, 2009. URL http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all&_r=0.

A. Kowarik and A. Meraner. *x12: x12 - wrapper function and structure for batch processing*, 2014a. URL http://CRAN.R-project.org/package=x12. R package version 1.5.0.

A. Kowarik and A. Meraner. *x12GUI: X12 - Graphical User Interface*, 2014b. URL http://CRAN.R-project.org/package=x12GUI. R package version 0.12.0.

A. Kowarik, B. Meindl, and M. Templ. *sparkTable: Sparklines and graphical tables for tex and html*, 2013a. R package version 0.9.7.

A. Kowarik, M. Templ, B. Meindl, and B. Fonteneau. *sdcMicroGUI: Graphical user interface for package sdcMicro*, 2013b. URL https://github.com/alexkowa/sdcMicroGUI. R package version 1.1.2.

B. Meindl. *sdcTable: Methods for statistical disclosure control in tabular data*, 2013. URL http://CRAN.R-project.org/package=sdcTable. R package version 0.10.3.

M. Templ, K. Hron, and P. Filzmoser. *robCompositions: An R-package for Robust Statistical Analysis of Compositional Data*, pages 341–355. John Wiley & Sons, Ltd, 2011a. ISBN 9781119976462. doi: 10.1002/9781119976462.ch25. URL `http://dx.doi.org/10.1002/9781119976462.ch25`.

M. Templ, A. Kowarik, and P. Filzmoser. Iterative stepwise regression imputation using standard and robust methods. *Comput Stat Data Anal*, 55 (10):2793–2806, 2011b.

M. Templ, A. Alfons, and P. Filzmoser. Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6 (1):29–47, 2012. doi: DOI:10.1007/s11634-011-0102-y.

M. Templ, A. Kowarik, and B. Meindl. *sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation*, 2014. URL `https://github.com/alexkowa/sdcMicro`. R package version 4.2.0.

M. van der Loo. The introduction and use of R software at Statistics Netherlands. In *Proceedings of the Third International Conference of Establishment Surveys (CD-ROM)*, Montréal, Canada, 2012. American Statistical Association. URL `http://www.amstat.org/meetings/ices/2012/papers/302187.pdf`.