Development and Current Practice in Using \textit{R} at Statistics Austria
Content

- Is \texttt{R} already state-of-the-art in the area of official statistics?
- Some benefits of using \texttt{R}
- \texttt{R} infrastructure at Statistics Austria
- \texttt{R} development at the methods unit at Statistics Austria. Few examples.
Using `R` is used in the production process from few countries.

A lot of new features for a lot of common tasks have been developed over the last years exclusively in R.

R replaces SAS (and SPSS) more and more. Ideal world: in 5 years nobody should use SAS anymore.

**Drawback**: (provocantly speaking) almost nobody can write good `R` code but they think they do.
Benefits 1/2

- R is a free and open-source environment
- R is a well-structured function- and object-oriented programming language
- Good tools for data manipulation
- Operators for calculations with vectors, matrices, arrays and tools for data analysis
- Excellent graphics
- Excellent features to produce dynamical reports for publication and web
is an environment, it is a mediator.

- interfaces to other programming languages such as C, C++, Java or Python
- 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 which allows much more freedom and flexibility compared to macro languages.
is an environment, it is a mediator.

- interfaces to other programming languages such as C, C++, Java or Python
- 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 which allows much more freedom and flexibility compared to macro languages.
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.
R in Statistics Austria
Infrastructure in Statistics Austria (short version)

- R is currently installed on more than 65 computers and on powerful virtual servers.
- R-administrators at Statistics Austria consists of three experts from the methods division.
- Contact person in every department.
- R-administrators are responsible for the container that includes RStudio, teaching material, documentation, examples.
- IT-department takes part of the container as installable file and distributes it.
- Wiki to share knowledge
- Teaching to increase knowledge
Teaching

- Two courses, basic and “advanced”
- content:
  - syntax and basics
  - data manipulation (also reshape2, plyr, data.table)
  - classes and object-orientation (mostly S3)
  - dynamic reporting (Sweave, knitr, markup, brew)
  - graphics (also ggplot2)
  - development, benchmarking, profiling, debugging, packaging
  - web-applications with R (shiny)
- ex-cathedra teaching is followed by exercises
Development
Our packages

- laeken
- (robCompositions)
- sparkTable
- sdcTable
- sdcMicro + sdcMicroGUI
- (TGUICore + TGUITeaching)
- VIM + VIMGUI
- x12

upcoming: CodaTable, simPopulation2, microSim
R-package laeken

- Standard and robust methods for estimating indicators from complex samples
- Variance estimation via calibrated bootstrap
- Provides a certain class structure to allow for easy handling of the functions and objects
R-package laeken

Documentation - type the following commands into R:

```r
install.packages('laeken')
vignette('laeken-intro')  ## JSS paper
vignette('laeken-standard')
vignette('laeken-pareto')
vignette('laeken-variance')
```

```r
library(laeken)
gini("eqIncome", weights = "rb050", data = eusilc)
```

Value:

[1] 26.48962

robust estimation, other indicators, variance, by breakdown is easy.
Breakdown by NUTS2.

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

Value:
[1] 26.48962

Value by stratum:

<table>
<thead>
<tr>
<th>stratum</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burgenland</td>
<td>32.05489</td>
</tr>
<tr>
<td>Carinthia</td>
<td>25.49448</td>
</tr>
<tr>
<td>Lower Austria</td>
<td>25.93737</td>
</tr>
<tr>
<td>Salzburg</td>
<td>25.01652</td>
</tr>
<tr>
<td>Styria</td>
<td>23.71190</td>
</tr>
<tr>
<td>Tyrol</td>
<td>25.24881</td>
</tr>
<tr>
<td>Upper Austria</td>
<td>25.49202</td>
</tr>
<tr>
<td>Vienna</td>
<td>28.94944</td>
</tr>
<tr>
<td>Vorarlberg</td>
<td>28.74120</td>
</tr>
</tbody>
</table>
GINI with EU-SILC data

Domain: Burgenland, female older than 65.
GINI, $k = 15$, Robust

Domain: Salzburg.
Variance. Naive Bootstrap.

```r
variance("eqIncome", weights = "rb050", design = "db040", breakdown="rb090", data = eusilc, indicator = a)
```

Value:
[1] 3.971415

Variance:
[1] 0.001700313

Confidence interval:

<table>
<thead>
<tr>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.879205</td>
<td>4.034146</td>
</tr>
</tbody>
</table>

Value by stratum:

<table>
<thead>
<tr>
<th>stratum</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1   male</td>
<td>3.796465</td>
</tr>
<tr>
<td>2   female</td>
<td>4.102162</td>
</tr>
</tbody>
</table>

Variance by stratum:

<table>
<thead>
<tr>
<th>stratum</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1   male</td>
<td>0.002834660</td>
</tr>
<tr>
<td>2   female</td>
<td>0.003648594</td>
</tr>
</tbody>
</table>

Confidence interval by stratum:

<table>
<thead>
<tr>
<th>stratum</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1   male</td>
<td>3.680585</td>
<td>3.899161</td>
</tr>
<tr>
<td>2   female</td>
<td>3.974129</td>
<td>4.208862</td>
</tr>
</tbody>
</table>
Use of small graphics (sparks) in text and graphical tables, like sparklines, boxplots, and bar charts. Finetuning is possible, e.g.

Graphical tables allowing to gain additional insights easily.
## League Table - Premier League 2011/12

<table>
<thead>
<tr>
<th>Team</th>
<th>Games</th>
<th>Performance</th>
<th>Wins</th>
<th>Draws</th>
<th>Losses</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man Utd</td>
<td>38</td>
<td>23</td>
<td>11</td>
<td>4</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Chelsea</td>
<td>38</td>
<td>21</td>
<td>8</td>
<td>9</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Man City</td>
<td>38</td>
<td>21</td>
<td>8</td>
<td>9</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Arsenal</td>
<td>38</td>
<td>19</td>
<td>11</td>
<td>8</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>Tottenham</td>
<td>38</td>
<td>16</td>
<td>14</td>
<td>8</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Liverpool</td>
<td>38</td>
<td>17</td>
<td>7</td>
<td>14</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>Everton</td>
<td>38</td>
<td>13</td>
<td>15</td>
<td>10</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>Fulham</td>
<td>38</td>
<td>11</td>
<td>16</td>
<td>11</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>A. Villa</td>
<td>38</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Sunderland</td>
<td>38</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>WBA</td>
<td>38</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Newcastle</td>
<td>38</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Stoke</td>
<td>38</td>
<td>13</td>
<td>7</td>
<td>18</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Bolton</td>
<td>38</td>
<td>12</td>
<td>10</td>
<td>16</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Blackburn</td>
<td>38</td>
<td>11</td>
<td>10</td>
<td>17</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Wigan</td>
<td>38</td>
<td>9</td>
<td>15</td>
<td>14</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Wolves</td>
<td>38</td>
<td>11</td>
<td>7</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Birmingham</td>
<td>38</td>
<td>8</td>
<td>15</td>
<td>15</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Blackpool</td>
<td>38</td>
<td>10</td>
<td>9</td>
<td>19</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>West Ham</td>
<td>38</td>
<td>7</td>
<td>12</td>
<td>19</td>
<td>33</td>
<td></td>
</tr>
</tbody>
</table>
sdcTable: Features

Main features of sdcTable

- Allows to protect tables having
  - multiple dimensions
  - arbitrarily complex structure of dimensions

- Standardized problem generation
  - Specification of dimensions
  - Data: micro-data, pre-aggregated data, ...

- Algorithms/Methods available:
  - primary cell suppression (different rules, custom)
  - secondary cell suppr. using a cut and branch algorithm (OPT, HITAS)
  - secondary cell suppr. using a (simple) variation of GHMITER
  - protection of 2 tables having common table cells
sdcMicro Package

- Includes (allmost) all popular methods for **microdata** protection.
- Is an user-friendly **highly object-oriented** implementation (only few commands must be known, such as `summary()`, `plot()`, etc.).
- Allows **reproducibility** and allows to apply anonymization methods in an **explanatory** manner.
- `sdcMicroGUI` for R users who still deny the benefits of command-line R.
- Package is used and distributed world-wide over Worldbank, OECD and IHSN.
highly optimised package

![Graph showing frequency estimation + risk measurement for different number of observations and time in seconds. The graph compares two methods: IHSN C++ and sdcMicro 4.1.0.]
VIM

(visualization and imputation of missing values)

- to explore and analyze the structure of missing or imputed values in data using graphical methods
- offers different built-in imputation methods to impute the missing values
- produce high-quality graphics for publications
- a simple graphical user interface (VIMGUI) allows an easy handling
VIM: Aggregation Graphic
VIM: Histogram

![Histogram Diagram]

Legend:
- Imputed
- Actual Data

Templ, Kowarik, Meindl (www.statistik.at)
VIM: Parallel Coordinate Plot
x12 Package

- provides an interface for the popular seasonal adjustment software X12ARIMA in R
- output from X12ARIMA is processed
- batch-mode for multiple time series
- parameters can be changed easily
- interactivity
- produces nice graphics for time series (with outliers, trend, ... )
x12: Time Series Plot with Outliers
x12: Forecast / Backcast

Time Series with Back− and Forecasts

<table>
<thead>
<tr>
<th>Date</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1948</td>
<td>100</td>
</tr>
<tr>
<td>1950</td>
<td>200</td>
</tr>
<tr>
<td>1952</td>
<td>300</td>
</tr>
<tr>
<td>1954</td>
<td>400</td>
</tr>
<tr>
<td>1956</td>
<td>500</td>
</tr>
<tr>
<td>1958</td>
<td>600</td>
</tr>
<tr>
<td>1960</td>
<td>700</td>
</tr>
<tr>
<td>1962</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- \( R \) is still the most powerful environment for statistics. Used since 2004 in partly at Statistics Austria.
- \( R \) is a mediator
- many solutions are available (add-on Packages in the area of official statistics)
- new strong focus on R in few departments at Statistics Austria. Teaching, infrastructure and the Wiki increase interest
- collaboration could be easy