Automatic data editing with open source R

Mark van der Loo, Edwin de Jonge

UNECE Work session on statistical data editing, 24-26 September, 2012
Statistical Value Chain

- Raw data
- Technically ok
- Clean microdata
- Aggregates
- Formatted output

Automatic data editing with open source R
Data editing process: design

Technically ok
Read data, rules
Ad hoc corrections
Typo's
Rounding errors
Error localization
Deductive imputation
Model imputation
Clean microdata

Automatic data editing with open source R
Data editing process: implementation

Control layer:
- Technically ok
- Read data, rules
- Ad-hoc corrections
- Typo's
- Rounding errors
- Error localization
- Deductive imputation
- Model imputation
- Clean microdata

Configuration layer:
- Control layer
  - R script
- Configuration layer
  - Correction rules
  - Edit rules
  - paths
- Workhorse functions
  - R-functions, bundled in a package

Automatic data editing with open source R
Data editing process: implementation

Automatic data editing with open source R
Example function: typo correction

```r
solveTypos <- function(E, dat, db, id){
  d <- correctTypos(E,dat)
  cors <- d$corrections
  storeLogRecords(db,
      id = dat[cors$row,id],
      variable = cors$variable,
      old = cors$old,
      new = cors$new,
      method = "typo"
  )
  d$corrected
}
```

- Each workhorse function is a simple wrapper around a library function (here: correctTypos of the deducorrect package).
- In-code documentation using Roxygen2 package (not shown).
Control layer: R script

```
dat <- detCorrections(corrections, dat, logdb,id="id")

dat <- solveTypos(rules, dat, logdb, id="id")

dat <- solveRounding(rules, dat, logdb, id="id")

el <- localizeErrors(rules,dat,verbose=TRUE)

dat <- deductiveImpute(rules, dat, el$adapt, logdb, id="id")

adapt <- is.na(dat)

dat <- impute(dat, logdb,id="id")

dat <- adaptValues(rules, dat, adapt=adapt, logdb, id="id")
```

The control layer is an R script with a few lines of preparations (reading data, config file), calling workhorse functions (shown), and writing.
Example of logging database

840 records of data with 42 variables, 42 editrules and 4 ad-hoc correction rules.

<table>
<thead>
<tr>
<th>id</th>
<th>variable</th>
<th>old</th>
<th>new</th>
<th>method</th>
<th>note</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LOONSOM121100</td>
<td>61.0</td>
<td>60.0</td>
<td>rounding</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:30:32</td>
</tr>
<tr>
<td>2</td>
<td>PERSONS110000</td>
<td>0.7</td>
<td>1.0</td>
<td>rounding</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:30:32</td>
</tr>
<tr>
<td>3</td>
<td>PERSONS110100</td>
<td>16.8</td>
<td>17.0</td>
<td>rounding</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:30:32</td>
</tr>
<tr>
<td>4</td>
<td>PERSONS110101</td>
<td>2.5</td>
<td>2.0</td>
<td>rounding</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:30:32</td>
</tr>
<tr>
<td>5</td>
<td>PERSONS110102</td>
<td>1.5</td>
<td>2.0</td>
<td>rounding</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:30:32</td>
</tr>
<tr>
<td>6</td>
<td>VOORZNG142000</td>
<td>NA</td>
<td>114.0</td>
<td>deductive</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:32:06</td>
</tr>
<tr>
<td>7</td>
<td>PERSONS110100</td>
<td>NA</td>
<td>21.1</td>
<td>regression</td>
<td>OPBRENG0000000</td>
<td>2012-09-05 11:32:07</td>
</tr>
<tr>
<td>8</td>
<td>PERSONS110100</td>
<td>21.1</td>
<td>1.0</td>
<td>adapt</td>
<td>&lt;NA&gt;</td>
<td>2012-09-05 11:32:09</td>
</tr>
</tbody>
</table>

The complete record history can be retrieved.
## Libraries used

<table>
<thead>
<tr>
<th>Method</th>
<th>library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-hoc corrections</td>
<td>deducorrect&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>typo corrections</td>
<td>deducorrect</td>
</tr>
<tr>
<td>rounding corrections</td>
<td>deducorrect</td>
</tr>
<tr>
<td>error localization</td>
<td>editrules&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>deductive imputation</td>
<td>deducorrect</td>
</tr>
<tr>
<td>regression imputation</td>
<td>default R</td>
</tr>
<tr>
<td>value adaption</td>
<td>rspa&lt;sup&gt;3&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

All packages available through CRAN.

<sup>1</sup>Van der Loo and De Jonge (2011-2012),  <sup>2</sup>De Jonge and Van der Loo (2011-2012),  <sup>3</sup>Van der Loo (2012).
Conclusion

- Built data editing street in short time using P&P functions from standard libraries.
- Delivers fully consistent data while retaining adapting history
- Documented as R-package using Roxygen2 package.
- Easily integratable with manual data editing software