An R-based data editing system

Mark van der Loo
R at Statistics Netherlands

- Strategic tool since 2010
  - Internal wiki, knowledge center, course.
- Used at:
  - National Accounts
  - Tourist statistics
  - Data collection with web robots (part of CPI)
  - Computing HSMR
  - Derivation of households
  - *Etc. etc. etc.*
- Used for:
  - (Complex) data manipulation
  - Analyses and regression
  - Visualisation
  - Data editing
Packages developed, at CRAN

- Data editing
  - editrules
  - deducorrect
  - rspa
- Data Visualisation
  - treemap
  - tabplot, tabplotd3
- Large data files
  - LaF (Large ASCII files)
Data editing packages

- **editrules**
  - Define rules
  - Verify data against them
  - Localize errors

- **deducorrect**
  - Deductive correction
  - Deductive imputation
  - Apply ‘direct rules’

- **rspa**
  - Adjust numerical records to satisfy rules

Rules defined with ‘editrules’ are reused by ‘deducorrect’ and ‘rspa’.
Automated data editing system for Child Care Centre Statistics

Prepare data, rule sets

.Net/SQL

Automated editing

.Net

Manual editing

.Net

.SPSS

.Child care statistics

Structural Business Survey
~800 records
~80 linear rules (balance edits)
~50 variables
Automated editing

- Correct with user-defined rules
- Thousand-errors
- Typos
- Roundings
- Error localisation with user-defined rules
- Error localisation based on F&H
- Deductive imputation
- NN imputation
- Adjust NN-imputed values

Raw data

if (x==0)
y<0

x+y == z
x > 0
x >= 0

Clean data

Log file
Example code: solve typing errors

```r
oplossenTikfouten <- function(E, dat, db, id){
  d <- correctTypos(E, dat)
  cors <- d$corrections
  opslaanLogRecords(db,
    id = dat[cors$row,id],
    variabele = cors$variable,
    oud = cors$old,
    nieuw = cors$new,
    methode = "tikfout"
  )
  d$corrected
}
```
# Results and process flow I: Cells

<table>
<thead>
<tr>
<th>Cells</th>
<th>Available</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Still available</td>
<td>Imputed</td>
</tr>
<tr>
<td>Available unadapted</td>
<td>Available adapted</td>
<td></td>
</tr>
</tbody>
</table>


Results and process flow I: Cells
Results and process flow II: Aggregates
**Results and process flow III: violations**

<table>
<thead>
<tr>
<th>Nr of checks: #Rules X #Records</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Violated</td>
</tr>
<tr>
<td>Still violated</td>
</tr>
<tr>
<td>Extra violated</td>
</tr>
</tbody>
</table>
Results and process flow III: violations
Results and process flow IV: measure of violation

An edit rule e can be understood as a 3-valued function of a record x:

\[
e(x) = \begin{cases} 
1 & \text{if } x \text{ satisfies } e \\
0 & \text{if } x \text{ violates } e \\
NA & \text{if } e(x) \text{ cannot be determined}
\end{cases}
\]

Tolerance: how much do I need to change x so e(x)=1?
Results and process flow IV: measure of violation (single rule)

An edit rule $e$ can be understood as a 3-valued function of a record $x$:

$$e(x) = \begin{cases} 
1 & \text{if } x \text{ satisfies } e \\
0 & \text{if } x \text{ violates } e \\
\text{NA} & \text{if } e(x) \text{ cannot be determined}
\end{cases}$$

**Tolerance:** how much do I need to change $x$ so $e(x)=1$?

Euclidean distance

In this case there is a closed-form solution
Results and process flow IV: measure of violation (single rule)

Positive tolerances per rule

Height of box ~ square root of nr of violations

Left axis denotes nr of unevaluated rules.
Results and process flow IV: measure of violation (multiple rules)

Given a set of rules $e_1, e_2, \ldots, e_n$ that a record $x$ must obey.

*How much do I need to change $x$, so that all $e_i(x) = 1$?*
Results and process flow IV: measure of violation (multiple rules)

Given a set of rules $e_1, e_2, \ldots, e_n$ that a record $x$ must obey.

How much do I need to change $x$, so that all $e_i(x) = 1$?

Euclidean distance

This can be computed with the rspa package.
Results and process flow IV: measure of violation (multiple rules)

Euclidean distance between actual and closest valid record.

A line traces one record.
Conclusions and outlook

- R-based, easy to build production-grade data editing system.
- Logging and indicators offer insight into
  - Quality of automated cleaning
  - Quality of data
- Future plans:
  - System is now being configured for another statistic
  - Implement general indicators (validator package)
  - Separate logging stream from data stream
- Reference
  - E. de Jonge and M. van der Loo *An introduction to data cleaning with R* (SN discussion paper nr 201313)