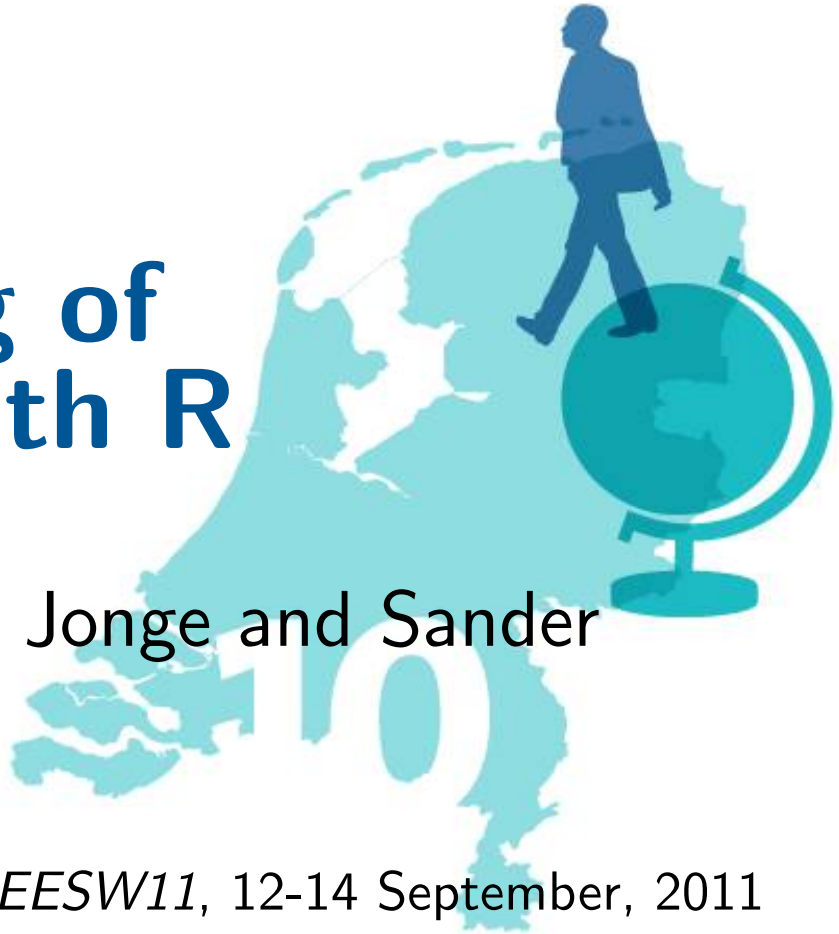
Automatic editing of numerical data with R



Mark van der Loo, Edwin de Jonge and Sander Scholtus

EESW11, 12-14 September, 2011

Introduction: R

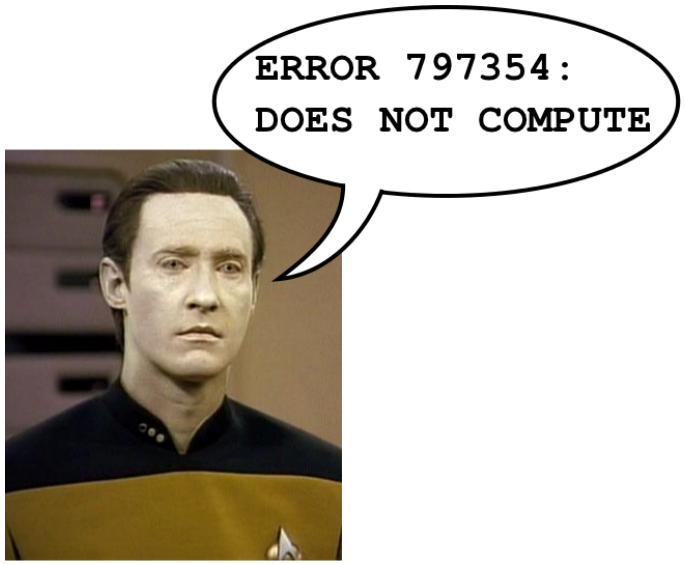
- Statistical computing environment and language
- World-wide standard
- Free as in speech and as in beer
- Modular: packages

Errors in data

cost	profit	turnover
12	342	8

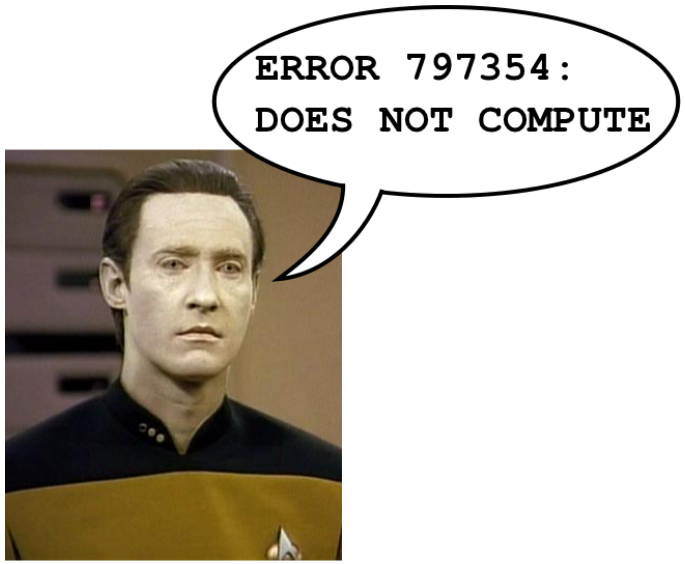
Errors in data

cost	profit	turnover
12	342	8



Errors in data

cost	profit	turnover
12	342	8

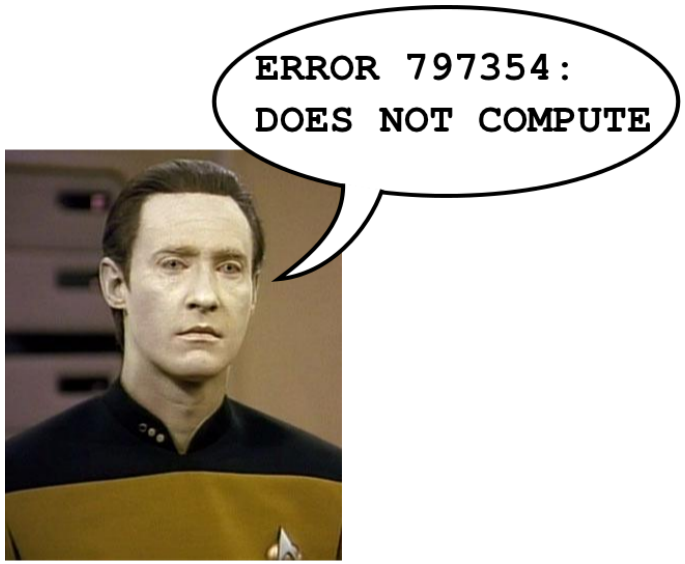


Expected:

$$\text{cost} + \text{profit} = \text{turnover}$$

Errors in data

cost	profit	turnover
12	342	8



Expected:

$$\text{cost} + \text{profit} = \text{turnover}$$

- Typical: $>$ 100 variables, several 100 rules, *e.g.* positivity, ratio's, account balances

Linear edit rules

Set of rules of the form

$$\mathbf{a}' \cdot \mathbf{x} \odot b$$

$$\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$$

$$\odot \in \{<, \leq, =, \geq, >\}$$

For example

$$\begin{aligned} \text{cost} + \text{profit} &= \text{turnover} \\ \frac{\text{profit}}{\text{turnover}} &\leq 0.6 \end{aligned}$$

Linear edit rules

Set of rules of the form

$$\mathbf{a}' \cdot \mathbf{x} \odot b$$

$$\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$$

$$\odot \in \{<, \leq, =, \geq, >\}$$

For example

$$\begin{aligned} \text{cost} + \text{profit} &= \text{turnover} \\ \frac{\text{profit}}{\text{turnover}} &\leq 0.6 \end{aligned}$$

With `editrules`:

```
> editmatrix(c(
+   "cost + profit == turnover",
+   "profit <= 0.6*turnover"))


Edit matrix:
   cost profit turnover Ops CONSTANT
e1    1      1     -1.0  ==        0
e2    0      1     -0.6  <=        0

Edit rules:
e1 : cost + profit == turnover
e2 : profit <= 0.6*turnover
```

Parsing and manipulation with `editrules`

Parsing:

- From text or `data.frame` to matrix: `editmatrix()`
- From matrix to text: `as.character()`

Checking:

- Check validity of records: `violatedEdits()`

Manipulation

- Check feasibility: `isFeasible()`
- Detect redundancy: `isObviouslyRedundant()`, `duplicated()`
- Find independent blocks: `blocks()`
- To echelon form: `echelon()`

De Jonge and Van der Loo (2011)

Errors with a clear cause and solution

	cost	profit	turnover
1	25	57	100
2	25	75	101
3	-25	75	100
4	25	100	75

Errors with a clear cause and solution

	cost	profit	turnover
1	25	57	100
2	25	75	101
3	-25	75	100
4	25	100	75

Errors with a clear cause and solution

	cost	profit	turnover
1	25	57	100
2	25	75	101
3	-25	75	100
4	25	100	75

The erroneous fields give a clue to the problem:

1. typing error
2. rounding error
3. sign error
4. value swap

. . .and therefore to their solution.

Deductive correction with `deducorrect`

- `correctTypos()`
  - Generate solutions to equality violations (QR decomposition)
  - Check if they correspond to typos, and if so: repair
- `correctSigns()`
  - Try combinations of sign flips and error swaps (binary tree)
  - Keep if it solves equality and inequality violations
- `correctRounding()`
  - Detect rounding errors
  - Draw a small, sufficient number of variables to change

All return a `deducorrect` object with

- Corrected data, status indator
- Corrections, logging info, timestamp

Algorithms by Scholtus (2008,2009) and Van der Loo, de Jonge and Scholtus (2011)

Errors with no clear cause

cost	profit	turnover
40	70	100

$$\begin{aligned} \text{turnover} &= \text{cost} + \text{profit} \\ \text{profit} &\leq 0.6 \cdot \text{turnover} \end{aligned}$$

Errors with no clear cause

cost	profit	turnover
40	70	100

$$\begin{aligned} \text{turnover} &= \text{cost} + \text{profit} \\ \text{profit} &\leq 0.6 \cdot \text{turnover} \end{aligned}$$

Errors with no clear cause

cost	profit	turnover
40	70	100

$$\begin{aligned} \text{turnover} &= \text{cost} + \text{profit} \\ \text{profit} &\leq 0.6 \cdot \text{turnover} \end{aligned}$$

Fellegi & Holt (1976): change the least (weighted) number of variables, such that every (derived) edit can be satisfied.

Error localization with `editrules`

```
> E <- editmatrix(c(
    "cost+profit==turnover",
    "profit <= 0.6*turnover"))

> record <- c(cost=40, profit=70, turnover=100)

> el <- errorLocalizer(E,record)

> el$searchBest()
$w
[1] 1

$adapt
    cost   profit turnover
   FALSE     TRUE    FALSE
```

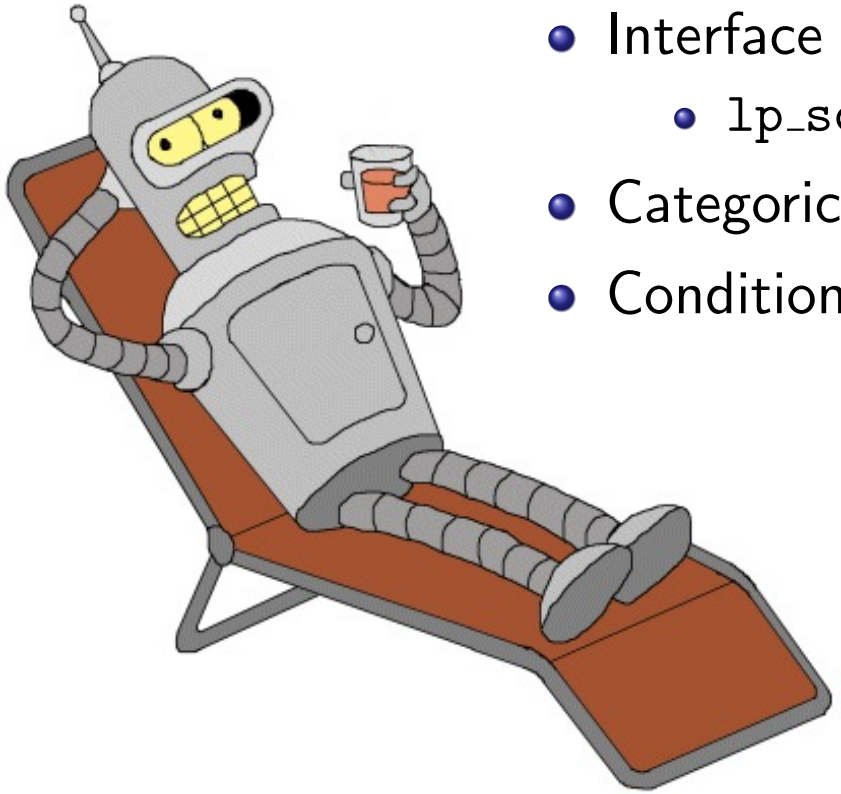
Algorithms by De Waal and Quere (2003), De Jonge and Van der Loo (2011)

`editrules` functionality

- `errorLocalizer()`
  - `searchNext()`, `searchBest()`, `searchAll()`
  - Control weights, search time, max weight, max adaptions
- `eliminate()`
  - Fast Fourier-Motzkin elimination in `editmatrix`
- `substValue()`
  - Substitute values in `editmatrix`
- `backtracker()`
  - Easy creation of binary tree search.

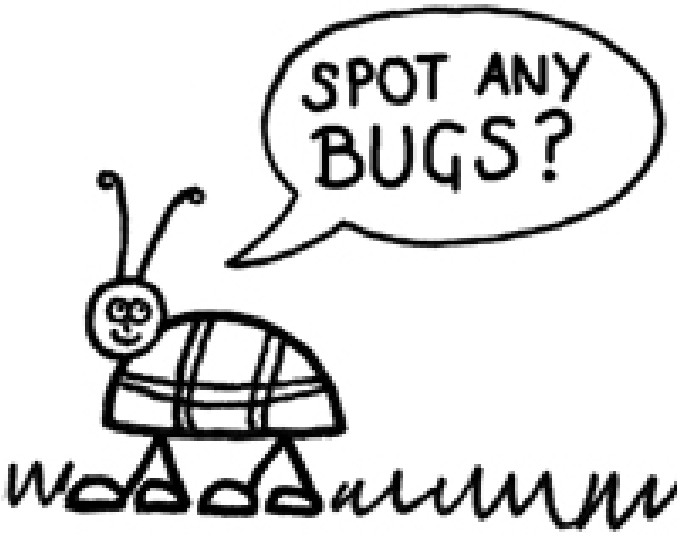
De Jonge and Van der Loo (2011)

The near future



- Interface with LP-solvers
  - `lp_solve`, `gplk`
- Categorical variables
- Conditional edits, mixed data

Thank you



m.vanderloo@cbs.nl
e.dejonge@cbs.nl

- De Jonge, E. and Van der Loo, M. (2011) Manipulation of linear edits and error localization with the editrules package. Discussion paper 201120, Statistics Netherlands The Hague/Heerlen
- De Waal, T. and Quere, R. (2003) A fast and simple algorithm for automatic data editing of mixed data. *Journal of Official Statistics* 19, 383-402
- Fellegi, I.P. and Holt, D. (1976). A systematic approach to automatic editing and imputation. *Journal of the American Statisticial Association* 71, 17-35
- Scholtus, S. (2008) Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data. Discussion paper 08015, Statistics Netherlands The Hague/Heerlen
- Scholtus, S. (2009) Automatic correction of simple typing errors in numerical data with balance edits. Discussion paper 09046, Statistics Netherlands The Hague/Heerlen
- Van der Loo, M. De Jonge, E. and Scholtus, S. Correction of rounding, typing and sign errors with the deducorrect package. Discussion paper 201119, Statistics Netherlands The Hague/Heerlen.