Proposal for useR!: 'READ.ISI'

Rense Nieuwenhuis*

March 27, 2008

1 Background

Due to technological and software development, it sometimes is no longer possible to automatically read older data-files into statistical software. Especially data-files that originate from the times magnetic tapes were used to store data are often distributed as raw (ASCII) data, without proper means to read those data into statistical packages.

However, for those interested in using data to perform longitudinal analyses, these older sets of data are very valuable.

In the Netherlands, the national archive for data storage (DANS) is currently organizing conferences on a unified and time-proof manner of storing data-files. But what to do with those data that already have become difficult to access?

2 The Problem

In a research project on fertility issues, it was found that the 'World Fertility Surveys'[1] are stored in a format that is no longer (directly) accessible to commonly used statistical software. Only data converted to ASCII directly from magnetic tape and a code-book are provided. The code-books are in a format specific by the 'International Statistical Institute' (ISI) and provides for each variable information on starting and ending positions in the data-file, value- and variable labels and information on missing values. However, no statistical software package presently used is known to be able to automatically read data based on this type of code-book.

It was required to read all variables into the statistical software manually. Variable names and value labels have to be assigned manually as well. This is not an inviting process and a highly laborious when many variables are needed.

*Author can be contacted at:
Email: contact@rensenieuwenhuis.nl
Telephone: +31 6 481 05 683

[1] http://opr.princeton.edu/archive/wfs/

3 The Solution

This problem may however be solved – in select cases – by using R-Project. Applying the flexible data-structure provided by R-Project, it was possible to read and interpret the code-books (meant for the human eye) and to use this to automatically read the data, add value and variable labels, assign missing values, and to do this for whole data-sets at once. The resulting syntax was transformed to the function called 'READ.ISI'.

```
V106      141  2   0   1   88      Remariee
                                      0    Non
                                      1    Oui
                                     88    Non rompue
V107      143  2   1   3   88   99 Etat actuel                    V104
```

Above, a small fragment of one of the code-books is shown. The function READ.ISI reads these fixed-width ASCII file twice. Once to read the variable names, labels, starting- and ending positions, and missing values (on the first and last row of the example above). The second time to read the value labels (in the middle rows of this example). As is illustrated on the last row of the fragment above, the value labels of variable 'V107' are identical to that of 'V104'. This is taken into account as well. Based on this automatic interpretation of this code-book, either the ASCII data-file is read, or a SPSS-syntax is created illustrating that people using other statistical packages can benefit from this function as well.

4 Proposal

Applicable to a select number of R users, but highly valuable for those who want to use (some) old data, this approach will help and inspire those who are interested in longitudinal analysis. Possibly, this approach can be transferred to the code-books of other collections of data. Therefore, I feel that this would make an excellent poster presentation on the userR! conference. On this poster the problem could be clearly illustrated and the steps needed to read this type of data automatically will be identified.