ffbase: statistical functions for large datasets
Edwin de Jonge¹, Jan Wijffels²

1. Statistics Netherlands, e.dejonge@cbs.nl
2. BNOSAC - Belgium Network of Open Source Analytical Consultants, jwijffels@bnosac.be

Keywords: Large datasets, memory constraints, modelling on large data

Statistical datasets used to be small, but nowadays it is not uncommon that the dataset is too large to be handled in R without encountering the frequently encountered Error: cannot allocate vector of size ... Mb issue.

To handle the R memory constraints, the ff package (Adler & Oehlschlägel et al.) was developed in 2008. It handles the memory constraint by storing data on disk. For day-to-day data munging, frequently used functionality from the base package had to be developed to make it more easy for an R developer to work with package ff. For this, the ffbase package has been developed to extend the ff package to allow basic statistical operations on large data frames, especially ffdf objects.

The ffbase package contains a lot of the functionality from the R’s base package for usage with large datasets through package ff. Namely

- Basic operations (c, unique, duplicated, fmatch, ffdfmatch, %in%, is.na, all, any, cut, ffwhich, ffappend, ffdfappend, rbind, fiftelse, ffseq, ffrep.int, ffseq.len)
- Standard operators (+, -, *, /, %, %, %, ==, =, <, <=, >, >=, &., |, !) working on ff vectors
- Math operators (abs, sign, sqrt, ceiling, floor, trunc, round, signif, log, log10, log2, log1p, exp, expm1, acos, acosh, asin, asinh, atan, atanh, cos, cosh, sin, sinh, tan, tanh, gamma, lgamma, digamma, trigamma)
- Selections & data manipulations (subset, transform, with, within, ffwhich)
- Summary statistics (sum, min, max, range, quantile, hist, binned_sum, binned_tabulate)
- Data transformations (cumsum, cumprod, cummin, cummax, table.ff, tabulate.ff, merge, ffdfdpoly, as.Date, format)
- Chunked functionalities (chunkify), writing & loading data (load.ffdf, save.ffdf, move.ffdf, laf_to_ffdf)

For modelling purposes, ffbase has bigglm.ffdf to allow to build generalized linear models easily on large data and can connect to the stream package for clustering & classification.

In the presentation, the ffbase package will be showcased to show that working with large datasets without having RAM issues in R is easy and natural for an R programmer.

References
