inca: an R Package for Integer Calibraiton

Luca Sartore[1,2], Kelly Toppin[2], Clifford Spiegelman[3]

[1]National Institute of Statistical Science (NISS)
[2]United States Department of Agriculture
National Agricultural Statistics Service (USDA NASS)
[3]Texas A&M University, College Park (TAMU)

lsartore@niss.org

Government Advances in Statistical Programming Workshop

October 24-25, 2018

Disclaimer

The Findings and Conclusions in This Preliminary Presentation Have Not Been Formally Disseminated by the U.S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy.

Presentation Outline

1. R package development
   - US Census of Agriculture
   - The calibration problem
   - Efficient computational solution

2. Steps towards publication
   - Programming and documenting the package
   - Debugging and revisions
     - Choice of license
     - Running conformity tests
     - Uploading the package in a repository

3. Concluding remarks

PART I

R PACKAGE DEVELOPMENT

- US Census of Agriculture
- The calibration problem
- Efficient computational solution

US Census of Agriculture

Every five years, USDA's National Agricultural Statistics Service (NASS) conducts the Census of Agriculture.

- The Census provides a detailed picture of U.S. farms, ranches and the people who operate them.
- It is the only source of uniform, comprehensive agricultural data for every state and county in the United States.
- NASS also obtains information on most commodities from administrative sources or surveys of non-farm populations (e.g. cotton ginning data).

Dual-System Estimation (DSE)

NASS uses DSE to adjust its estimates by generating weights assigned to each data-record.

- DSE requires two independent surveys to produce adjusted estimates for under-coverage, non-response and incorrect farm-classification at the national, state and county levels.
- The adjusted weights are used as starting values for the calibration process.
- The weights are calibrated to ensure that the Census estimates are consistent across all levels of aggregation and in agreement with information from other sources.

All details can be found in Young et al. (2017).

Calibration

A solution $\hat{w}$ such that $T = Aw$, where

- T is a vector partitioned into y known and y^* unknown population totals,
- A is the matrix of collected data from a population, and
- w is a vector of unknown weights.

Calibration finds the solution of the linear system $y = \tilde{A}w$, where

- $\tilde{A}$ is a sub-matrix of the collected data.

NASS publishes its estimates by using
integer weights
to avoid fractional farms.

The solution

Rounding occurs before calibration, and successively, the weights are adjusted through a discrete coordinate descent algorithm (Sartore et al., 2018).

Description of the algorithm

1. All unfeasible weights are truncated to the closest feasible boundary, and in order to minimize the objective function, non-integer weights are then rounded sequentially according to an importance index based on the a gradient.
2. Each weight, according to the magnitude of the gradient, is allowed to move in unit-shifts which decreases the objective function.

PART II

STEPS TOWARDS PUBLICATION

- Programming and documenting the package
- Debugging and revisions
  - Choice of license
  - Running conformity tests
  - Uploading the package in a repository

Programming environments

Distinction between INCA, as a production software at NASS, and the R package on CRAN (Sartore and Toppin, 2016).

NASS	**CRAN**
• Developed in C • Interface with SAS • Two distinct objective functions • Stand-alone library	• Developed in C++ • Interfaced with R • Several objective functions available to the user • Requires other R packages for dependencies

The package was programmed and tested on a SuSE Linux Enterprise Server as a prototype for the production software in a unix environment.

Debugging

Data from 2012 US Census of Agriculture were used to debug the code and introduce some computational tricks (e.g. the use of sparse matrix representation).

Each state was processed separately by allowing for parallel computing at a specific stratification level (Sartore et al., 2018) and reducing computational time from 3 days to 20 minutes.

Extensively tests were performed to study the ability to produce reliable estimates with collinear data, misspecified targets, and different sets of DSE weights.

For an example, see
https://www.census.gov/fedcasic/fc2016/ppt/1_5_Integer.pdf

Internal revisions

Once the main routines were developed, two other expert programmers reviewed the code to assure that the results are not affected by record permutations or processor type (e.g. i386, x86_64).

The convergence properties of the algorithm were also studied. INCA algorithm locally converges to an integer solution at least linearly.

The documentation of the package was written, then it was reviewed by our deputy director. Some examples with few random numbers were also provided.

Licensing

Federal agencies have policies and guideline related to open-source developed and distribution.

- For general regulations on open-source software development, see https://sourcecode.cio.gov/.
- For useful information on the choice of license, see https://github.com/18F/open-source-policy/blob/master/policy.mdv/styleguide/
  US public domain software uses Creative Commons Zero (CC0) to waive copyright internationally.
- For other details, see https://dodcio.defense.gov/open-source-software-faq/.

CRAN policies

After complying with the US federal laws, and before submitting the package to the CRAN, the following steps are required (https://cran.r-project.org/web/packages/submission_checklist.html):

1. Conform the package also to CRAN policies (https://cran.r-project.org/web/packages/policies.html).
2. Check the package and its documentation on several platforms (https://github.com/r-hub/rhub).
3. Fix all errors and warnings.
4. Repeat step 2 and 3 until no errors and warnings are returned.

Uploading the package

The current package upload procedure has been automated by the introduction of a web-server (https://cran.r-project.org/submit.html).

The submission consists of three steps:

1. The upload of the package and the maintainer information.
2. Submission to CRAN members for further tests before publication.
3. Receiving the confirmation that the package is online.

Maintenance

Updating the R package:

- New function, routines, and algorithms can be added to an existing R package.
- Changes within R make the package obsolete (and it could be removed from CRAN).
- Policies can change over time and new rules need to taken into consideration.

New tools and guidelines will be developed to simplify the following steps:

- Programming (e.g. Rstudio, Qt Creator, Visual Studio Code, Eclipse IDE).
- Debugging (e.g. R-hub, valgrind).
- Documenting (e.g. Roxygen2) and tracking changes (e.g. Git, Mercurial).

PART III

CONCLUDING REMARKS

Concluding remarks

- Publishing an R package on CRAN can be tedious process, but it guaranties some standards when developing open-source software within the federal government.
- Having an R package on CRAN forces the maintainer to keep the package up to date with changing technology.
- Improve the transparency of the statistical methods adopted to make analyses, or even produce official estimates (e.g. via MS R Open, which is becoming a valid alternative to SAS).
- R packages developed within the federal government should be published on CRAN and successively on `https://code.gov`.

Selected References

Sartore, L. and Toppin, K. (2016). *inca: Integer Calibration*. R package version 0.0.2.

Sartore, L., Toppin, K., Young, L., and Spiegelman, C. (2018). Developing integer calibration weights for Census of Agriculture. *Journal of Agricultural, Biological and Environmental Statistics*, Accepted.

Scott, T. and Rung, A. E. (2016). Memorandum for the heads of departments and agencies.

Wickham, H. (2015). *R packages: organize, test, document, and share your code*. O'Reilly Media, Inc.

Young, L. J., Lamas, A. C., and Abreu, D. A. (2017). The 2012 Census of Agriculture: a capture–recapture analysis. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):523–539.

Thank you!

Questions?

Luca Sartore, PhD `lsartore@niss.org`
Kelly Toppin, PhD `kelly.toppin@nass.usda.gov`
Clifford Spiegelman, PhD `cliff@stat.tamu.edu`