Using *R* for official statistics: Census of Foreign Capital in Brazil.

Carlos Cinelli[1,2,*] , Rodrigo Wang[1]

1. Brazilian Central Bank
2. University of Brasilia
*Contact author: carlos.cinelli@bcb.gov.br

Keywords: official statistics, data imputation, census of foreign capital in brazil, international investment position.

The Census of Foreign Capital in Brazil (Census) has been carried out quinquennially since 1996 by the Brazilian Central Bank (BCB). Its major purpose is to measure the stock of Foreign Direct Investments in Brazil (FDI), necessary to compile the International Investment Position (IIP) statistics. As of 2011, the Census was split into two surveys: the 5-year Census and the Annual Census, the latter targeted to large enterprises only. The distribution of the FDI stock is heavy tailed, so large enterprises comprise 80-90% of the total value, but represent only a small fraction (10-20%) of the total number of respondents, thus reducing the cost of the survey without much loss of information.

As for the data of the missing respondents, we mostly replicate their latest survey values and add their flows registered in the Balance of Payments. But this ***is not*** as trivial as it sounds. During the 5-year gap between the complete Censuses a lot can happen: new companies start, some companies will close, other companies will merge with each other, some companies will not have foreign investors anymore or the nationality of the foreign investors may change. To deal with these facts we must gather information from other sources (for example, the Brazilian IRS) and perform some data analysis in order to decide which data will or will not be imputed, and how it will be imputed. The further you are from the latest 5-year Census, the more complicated this task gets.

Microsoft Excel is widely used for data manipulation and data analysis in Central Banks and International Organizations (like the IMF). So, unfortunately, our first choice was to use Excel. This was prone to a lot of operational errors - *Reinhart-Rogoff style*. It required the use of many different spreadsheets and files: a cumbersome process to manage that was hard to find a bug when there was one. It was not easy to immediately reproduce the results, and it was not really clear to an outsider where and when data modification was taking place, because of all the cross-references between worksheets. So we decided to create *R* packages with functions that automate the process.

We have developed three packages (names in Portuguese): **censo.criar.base,** which gathers and combines all information necessary to the imputation and compilation into a new database; **censo.extrapolar,** which automates the imputation; and, **censo.quadros,** with functions to calculate the main statistics and analysis. The publication's process time *reduced from 1-2 weeks to a couple calls in the command line* (that takes only a few minutes). It is now easier to track bugs and errors, because all data comparisons and transformations are clearly stated on the codes. Another advantage is that new kinds of data exploration and visualization, that once were not possible, are now easily available through *R*. This has helped the development of a more structured ***validation*** and ***exploration*** of the data – and those are the new packages we are working on.

The goal of our presentation is to describe the imputation/validation/publication process of the FDI statistic in BCB, focusing on the application of *R* in our workflow and deepening the discussion of the main advantages/disadvantages of using *R* for official statistics.