Paper size: A4, character: Times New Roman, all margins 2 cm

ICAS VI
SIXTH INTERNATIONAL CONFERENCE
ON AGRICULTURAL STATISTICS 2013

Title of the Paper

Filippo Gheri, filippo. gheri@fao.org
Michael Kao, michael.kao@fao.org
Amy Heyman, amy.heyman@fao.org
Food and Agriculture Organization of the United Nations
via delle Terme di Caracalla
Rome Italy

ABSTRACT (all Caps, character 14 pt, bold, adjust Left)

Compiling hundreds of statistics from different sources with a traditional approach, such as manual preparation of Excel tables, can be very labour intensive and prone to error. Furthermore, knowledge and expertise is difficult to transmit, thus resulting in inconsistent results and treatment over time. Therefore the Statistics Division of FAO implemented the use R and LaTeX as the new architecture for a sustainable and cost-effective way to produce its Statistical Yearbook. The R packages include all of the steps of the process: data retrieval, merging data, computations, and creation of visualizations. On the other side, LaTeX provides the layout structure. Because all steps are well documented, this approach increases the longevity and coherence of the publication. The combined power of R and LaTeX makes this a new data publication in line with the open data philosophy. And the use of open source software and the availability of the package – therefore total transparency – makes the entire procedure available to anybody.

Keywords: Keyword 1, Keyword 2, Keyword 3 (character 12 pt, double-justified)

1. Title of the Section (character 14 Pt, bold)

1.1 The FAO Statistical Yearbook series

The FAO Statistical Yearbook series[1] started in 2004, consolidating and replacing four previous FAO publications – the FAO Bulletin of Statistics, and the FAO Production, Trade and Fertilizer Yearbooks. The purpose was to provide the users with a selection of indicators on food and agriculture by country. The data were taken from FAOSTAT[2], the Organization’s corporate statistical database, as well as from several other FAO divisions and sources, mainly within the UN system.

[1] http://www.fao.org/economic/ess/ess-publications/ess-yearbook/en/#.UbHn8Nhj7To
[2] http://faostat3.fao.org/home/index.html

Paper size: A4, character: Times New Roman, all margins 2 cm

In 2012, the Statistics Division launched a Statistical Yearbook (SYB) with a very new look and feel. The publication now presents a visual synthesis of the major trends and factors shaping the global food and agricultural landscape and their interplay with broader environmental, social and economic dimensions. In doing so, it strives to serve as a unique reference point on the state of world food and agriculture for policy-makers, donor agencies, researchers and analysts as well as the general public. Thus far, two global and two regional Yearbooks have been released. A Pocketbook, which is a reduced version of the Yearbook but presented instead by country, has also been designed.

The novelty of the Yearbook is not only the content, which fills a global gap of presenting up-to-date information on agriculture and food security, but also the efficient method of production, which sets a precedent for dissemination. The publication is generated with the statistical software R[3] in combination with the typesetting program LaTeX[4].And both are open source softwares. The two programs work in concert to provide a seamless process in generating data publications and therefore more efficient dissemination tools. The way of working within the team has also changed. The use of open-source software like GitHub[5], SVN[6] and Dropbox[7] provides continuous tracking of all changes to documents and to be able to simultaneously work on documents.

1.2 Motivation of using an open-source programming language

The initiative of using an open-source programming language for the SYB was taken for two specific reasons: one related to open-source software and the other to the use of a programming language. Following the GNU free software definition[8], "'free software' means software that respects users' freedom and community. Roughly, the user have the freedom to run, copy, distribute, study, change, and improve the software. With these freedoms, the users (both individually and collectively) control the program and what it does for them". This means that moving from proprietary software to open-source software first reduces the increasing license costs. It also makes the process completely transparent, which is particularly important because it allows other researchers to understand how the program works, giving them the possibility to improve on the work.

The second reason concerns the use of a programming language. Among the advantages, the most important is reproducibility, for which access to the code is a precondition. Reproducibility is the norm in the academic scientific field. This property allows one to verify, improve and reproduce research for future use. This has tremendous potential for the user and, in this case, for FAO. Furthermore, since the methodology and codes are available publicly, the active community can also provide extremely valuable feedback to improve the overall system.

Other advantages involve automation, the text based source, and communication. The high level of automation enormously reduces the manual operations prone to errors and saves significant time in updating data. Moreover, the fact that a programming language is essentially text makes it possible to easily keep track of all of the changes through version/revision control systems, to simply print out the code and store it in the long term, and to copy and paste sections that could be reused for other purposes. Finally, code permits reproducible examples so that problems can then be shared within the community, therefore increasing the chances to solve issues and, at the same time, providing assistance to other users.

Therefore, this approach increases the sustainability, transparency and coherence of the publication because all data manipulation and treatment are recorded within the code. Transparency in this type of publication is crucial. The philosophy of the new SYB, which publishes country-level data and

[3] http://www.r-project.org/

[4] http://www.latex-project.org/

[5] https://github.com/

[6] http://subversion.tigris.org/

[7] https://www.dropbox.com/

[8] https://www.gnu.org/philosophy/free-sw.html

aggregates, should be examined under the same standard. It is important to publish these figures as well as the methodology used to prepare the data. This type of feedback can also lead to a greater harmonized framework.

The major downside, however, of using a programming language is the steep learning curve. In addition, the speed is sometimes inefficient, and slow, especially for exceptionally large datasets. And, given its data analysis focus, the programming infrastructure is not very well developed. As a result, the benefits of this tool applies primarily to those who work on it on a daily basis.

1.3 Motivation of using R and LaTeX

R is usually defined as a free software environment for statistical computing and graphics. However, rather than a domain specific language for statistics, R is more precisely a domain specific language with broad functionality. As an open-source software, R can be used for a variety of purposes. The strong R community has developed more than 4,000 documented packages that span several fields of science. The user dialogue is extremely active and works to solve problems within the community. The freedom to study how the specific functions work allows the user to adapt the algorithms to their specific needs and therefore continuously improve packages. Last but not least, the high connectivity to other languages and software makes R a flexible tool that is able to build almost anything users need. In the SYB process, R is used for data processing, analysis and dissemination. More specifically, through R we automatically download variables, import datasets, merge together datasets with different standards, construct new variables, compute aggregations, and finally disseminate data through charts, tables, and maps. At the end of this process, the dissemination objects are automatically translated in LaTeX code.

LaTeX is an open-source high-quality typesetting systems that includes features designed for the production of technical and scientific documentation. Together with R, LaTeX is open source with no licence costs. Moreover, LaTeX is a markup language and therefore has many advantages with the programming language category. Compared to other word processing programs like MS Word or LibreOffice Writer, LaTeX accommodates mathematical notation easily, controls sectioning, cross-references, tables and figures, and automatically generates bibliographies and indexes. It separates content and style, and therefore the user can write the document without also changing format. It allows multi-lingual typesetting and guarantees a perfect consistency throughout the book (e.g. in the management of the colours), thus facilitating enormously the printing step without intervention from graphic designers. And finally, LaTeX also moves seamlessly from print to web version in one click.

LaTeX is used to typeset the entire SYB publication. The dissemination objects translated in LaTeX code by R are automatically included and formatted within the publication through the specific class *faoyearbook*. This package defines all the needed commands to delineate the structure and build the publication. The package is geared towards teh Yearbook, but can bea adapted easily to create other publications. This makes the process completely exportable/applicable with a relatively small amount of time needed to design the book layout.

1.4 Revision control systems

A revision control system is a way of managing changes to files. This software is essential for developers and, although they differ in terms of functionalities, they usually have similar advantages. First of all, a revision control system allows users to revert to previous versions of a document and to track all the modifications within the file. Secondly, users can work simultaneously on the same files on multiple computers, even if people are offline. The merging operation solves possible conflicts or, if this operation is not available, the file locking method prevents concurrent access simply by locking the specific file. Third, it gives the possibility of creating branches of a project and identifying a snapshot of this through tags.

It is usually very difficult to manage binary files (e.g. excel spreadsheet) with revision control systems, and it is even more difficult to run some basic operations, like merging, in case a conflict arises. R and LaTeX are essentially scripts. This means that all of the changes are automatically tracked and documented and that everybody in the team is informed about activity. Moreover, the possibility of creating different branches for a single project leads to one stable version. This is crucial to monitor the development of the project as well as, on the practical side, to keep separate what has been already tested from what is under development. In the SYB process, Dropbox, SVN, and GitHub are used for different purposes. We store in Dropbox all those files that require an immediate update and that are usually used by just one member of the team. GitHub stores the R packages that are publicly available, while SVN is used for the LaTeX package and the final output that is not publicly available.

2. SYB production

The construction of the SYB revolves mainly around processing, analysis and dissemination steps as described by the Generic Statistical Business Process Model[9]. Processing and analysis are entirely done through R, while data dissemination occurs with both R and LaTex. More specifically, two R packages have been created. The FAOSTAT package hosts a list of functions to download, manipulate, construct and aggregate agricultural statistics, while the FAOSYB package includes functions to disseminate these statistics.

2.1 Data retrieval

Data retrieval involves all operations to import variables from different databases. The simplest way is by querying a database. When this option is not possible, the user has to manually import the necessary indicators. The FAOSTAT package provides the functions to automatically download variables from the FAOSTAT[10] and the World Data Bank[11] databases as well as the algorithms needed to manually import specific sets of data.

2.1.1 Download data from FAOSTAT and World Data Bank

The function *getFAOtoSYB* collects data using *getFAO* and processes these in order to retrieve the dataset in an easily manageable format. *getFAO* allows to access to FAOSTAT data through the FAOSTAT API. The user is facilitated in the construction of the API by the function *FAOsearch*, which, with very few steps, provides users with the needed codes to build the query. Analogous is the behave of *getWDI* and *getWDItoSYB* functions to download the World Development Indicators.

2.1.2 Import data manually

Importing datasets with different structures and *ad hoc* variables in various formats requires a flexible programming language able to dialogue with different software. Nevertheless, in most cases, this is not enough. First, organizations often have different constructions of the world, and do not have a common classification system and are not necessarily harmonized across countries and borders., Therefore all of the (frequent) exceptions made by organization to the reference country classification system should bestudied on a regular basis. . Only then is it possible to accurately design the software to address these issues.

Aggregates can therefore not be imported unless the definitions match exactly, which often is not the case. In the Yearbook production, these aggregates are not discarded but are used to check the

[9] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model.

[10] http://faostat3.fao.org/home/index.html.

[11] http://databank.worldbank.org/data/views/variableSelection/selectvariables.aspx?source=world-development-indicators.

aggregates we compute computed for potential errors in the aggregation methodology and/or weighting variables.

Three are the main issues attempting to combine different country classification systems. The first is differences in the country definitions, which aer generally due to varying legal recognitions. The second is changes in the country composition over time. China is one such example. The World Bank disseminates China (excluding Taiwan), Hong Kong and Macao. On the contrary, FAOSTAT disseminates China (including Taiwan, Hong Kong, and Macao), while other sources disseminate China (including Taiwan), Hong Kong, and Macao. It is clear that in the three cases "China" represents three different realities. The second aspect then complicates this problem by adding the time dimension. South Sudan was recognized by the United Nations on the 9 July 2011. However, statistics reported by the Sudan in the same year can also include data for South Sudan, which leads to double counting. Finally, the analysis is further complicated by disputed territories and economic unions, such as Ilemi Triangle and Belgium-Luxemburg, which do not have representation under most of the country coding systems.

A precise matching is thus essential in order to merge. The SYB uses the M49 country classification system following the idea of a needed convergence to a common international standard. However, for internal purposes the FAOSTAT country coding system is used. The reason is that this coding system is disaggregated enough to allow maximized matching. The functions *fillCountryCode* and *translateCountryCode* provided by the FAOSTAT package help respectively in filling and translating the country code when just country names are provided . Nevertheless, a perfect matiching is not always possible. If the country classification used considers the dissemination of China (including Taiwan), Hong Kong, and Macao, but the time series includes just China (including Taiwan, Hong Kong, and Macao), then it is impossible to disaggregate the data, and a footnote is needed.

2.2 Merging datasets

Merging is a typical data manipulation step in daily work – albeit non-trivial – especially when working with different data sources. Within the FAOSTAT package, the built-in *mergeSYB* merges data from different sources as long as the country coding system is identified. A precondition for this operation is the correct structure of the manually imported datasets following the rules described in the previous paragraph.

2.3 Scaling

In theory, data should be processed and stored in the base unit (e.g. kilograms) and, if needed, disseminated with an attached multiplier (e.g. thousand kilograms) or with a different scale of the same measure (e.g. tonnes). Nevertheless, very often this is not the case and, as a result, they need to be rescaled. Both of these operations are done with functions external to the FAOSTAT package because of the way of treating measurement units by different users, both in terms of different coding/naming system used and results to be obtained. This heterogeneity would imply a complex matching of different systems that, for the moment, has not been developed.

2.4 Construction of new variables

The FAOSTAT package can automatically construct new variables, including growth rates, shares, indices and relative changes. Two types of shares can be computed. If just one variable is used, the "share of total" option checks the weight of a specific country/aggregate vis-a-vis the total. There are also two types of growth rates within the package: the least squares growth rate and the geometric growth rate. The least squares growth rate is used when the time series is of sufficient length. The default is at least five useable observations. However, if the time series is sparse and more than 50 percent of the data are missing, then the robust regression is used. Furthermore, for a specific time series both index number and relative change can be computed. The first one need the base year, while the second the year interval.

Paper size: A4, character: Times New Roman, all margins 2 cm

2.5 Aggregation

Aggregation is another data manipulation step that is commonly over seen. In most of cases the already computed aggregates cannot be used because: a) a different country classification system is used; b) the change in the country composition over time is treated differently. For the same reasons, it is difficult to get exactly the same results, and address missing values in a harmonized manner.

At this point of the process, the user must a) converge with a specific country classification system, and b) generate consistent and comparable aggregates. In the specific case of the SYB, M49 list[12] is used. This problem has been addressed in the SYB by a two steps aggregation process.

2.5.1 Country level

The starting point is a set of countries that in our case is as disaggregated as possible through the FAOSTAT coding system. The objective is then to match the information with the M49 country level definition. In other words, if the big set of countries is composed by China, Taiwan, Macao, Hong Kong, Tanzania and Zanzibar, the objective is to get a set of countries composed just by China (including Taiwan), Macau, Hong Kong, and Tanzania. In the first aggregation step, we merge together the countries that go together following the M49 system. In the example, we merge together China and Taiwan, and Tanzania and Zanzibar. The *aggCountry* function allows to aggregate territory entries into country or higher level classification based on the relationship specified. It is clearly important to specify the correct method and weighting variable, otherwise the aggregate will produce erroneous output. Nevertheless, no aggregation rules are applied at this step.

2.5.2 Geographic and economic level

In order to compute geographical, economic, and political aggregates a further aggregation step is needed. The first problem is that a hierarchical approach cannot be used. It would be a logical first step to compute the aggregations starting from the set of countries obtained after the first aggregation. However, this cannot be done for two reasons. First, while the first step exclusively follows a political criterion, the second could follow other rules, such as geography. External territories are a perfect example. They could politically belong to countries that are geographically on the other side of the globe. These territories are incorporated into the “mother” countries in the first step, but they do not follow the “mother” country in the second. The first implication is that the countries of a specific region would therefore not sum up to the regional aggregate.

Second, following a hierarchical approach would mean not considering the time dimension, thus the historical evolution of the country composition. This would violate the comparability over time. Clearly, the final country list reflects the last updated world composition. This does not create challenges for current year. However, given that we are interested in computing aggregates for the past years, we need to consider how the world composition has changed over time. If we want to compute an aggregate for Africa in 2013 we need to include South Sudan and Sudan. On the other hand, this rule is not valid in 2010 when South Sudan and Sudan were a single country.. For this reason, the Africa aggregate can only be computed consistently by including South Sudan, Sudan and Sudan (former). And therefore the hierarchical approach does not apply. The strong assumption that should be verified is that data are mutually exclusive. The presence of data for all the three Sudans would imply a double counting.

This means that the second aggregation step starts again from the big set of countries and needs a partially different relationship. In this case further specifications are also needed in order to address the problem of the missing values, which can render the aggregates incomparable. Two rules are implemented to ensure the aggregates computed are meaningful and comparable: first, a minimum threshold (default 65 percent) in which data must be present; and second, the number of reporting entities must be similar over the years (default tolerance is 15) because it does not make sense to

[12] http://unstats.un.org/unsd/methods/m49/m49.htm

Paper size: A4, character: Times New Roman, all margins 2 cm

compare aggregates for 1995 and 2000 if the number of reporting countries vastly differ. Both these rules are applied automatically by the function *aggRegion*.

2.6 Analysis

Exploratory data analysis is fundamental before conducting any modelling operations and data dissemination. In a publication, this type of analysis is crucial in order to understand the main messages behind the data and to decide the central idea to be passed to the user through a specific object, sub-section and section. Itshould help clarify what a specific dissemination object is trying to communicate, how this message fits within the sub-section idea and how it is linked to the other messages.

Dealing with international datasets means being aware that, in most of cases, the aggregates we would like to show are unavailable due of data sparsity. The *sparsityHeatMap* function provided by the FAOSTAT package checks data sparsity for all variables, across country and time. the function generates a plot grouped into four panels. The first three panels group the country by their of value, while the last shows countries with no values.

Another tool within the FAOSTAT package is the *tsPanel*. The advantage of the plot generated by this function is to identify the behaviour of a specific variable, in particular if one was to build models or carry out imputation. The characteristics that govern the variable and the transferability of country information determines what type of model is available.

2.7 Dissemination

While the R FAOSTAT package focuses on data processing and analysis, the R FAOSYB package supports the user in the dissemination phase. The functions *theme_syb* and *plot_color* define a style and a set of colours to be applied across the publication in order to ensure consistency across the book. *plot_data* and *plot_dictionary* help the user to create predetermined types of charts that come from the R package *ggplot2*. Furthermore, the functions *GAULspatialPolygon*, *map_breaks* and *plot_map* help use maps in *ggplot2* and the shape files provided by the GAUL project[13]. Tables are generated using internal codes have not yet been added to the package due to their complexity. However, what is important is that charts, maps, tables and mini tables are essentially R code, and, for this reason are easily reproducible and updatable. In the end, these objects, captions, sources and metadata are automatically translated into LaTeX code

2.8 Typesetting

The typesetting of the publication is then entirely done with LaTeX. Dissemination objects, captions, sources, text, bullet points and metadata are assembled together by the SYB specific class *faoyearbook*. Automatically, LaTeX controls sectioning, cross-references, and indexes. The bibliography is done with BibTeX[14] and it is read automatically by LaTeX.

2. Conclusion

The new FAO Statistical Yearbook

REFERENCES (all Caps, character 14 pt, bold, adjust Left)

Berger J. (1990) Robust Bayesian analysis: sensitivity to prior, *Journal Statistical Planning and Inference*, 25, 303-328.

[13] http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691.

[14] http://www.bibtex.org/.

Paper size: A4, character: Times New Roman, all margins 2 cm

Cooper M. C., Milligan, G. W. (1988) The effect of measurement error on determining the number of clusters in cluster analysis, in: *Data, Expert Knowledge and Decision*, Gaul, W. & Shader, M. (Eds.), Springer, 3 19-328.

References (text): Line Spacing Single, Adjusted Left. All references should be given in alphabetical order. References should be indicated in the text by name(s) of the author(s) and the year of publication in round parentheses.