Efficiency and Effectiveness: The FAO Statistical Yearbook

Amy Heyman

Food and Agricultural Organization of the United Nations

amy.heyman@fao.org

Markus Kainu

Food and Agricultural Organization of the United Nations

markus.kainu@fao.org

Filippo Gheri

Food and Agricultural Organization of the United Nations

filippo.gheri@fao.org

April 15, 2015

Abstract

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 LaTeXas 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: retrieving and merging data, conducting computations, and creating visualizations. On the other side, LaTeXprovides 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 LaTeXmakes this a new data publication in line with the open data and open science philosophies. And the use of open source software and the availability of the package – therefore total transparency – makes the entire procedure – and the possibility of using the technology to produce other publications – available to anybody

Contents

1 Overview

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. In 2012, the Statistics Division launched a new Statistical Yearbook (SYB), which presented a very new look and feel. The products that are a part of this suite of publications 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, they strive to serve as a one-stop-shop on the state of world food and agriculture for policy-makers, donor agencies, researchers and analysts as well as the general public.

```
library(ggplot2)
library(dplyr)
df <- mtcars
dat <- df %>% group_by(carb) %>% dplyr::summarise(horsepower = mean(hp, na.rm=TRUE))

# generated object
caption_text <- "Number of carburetors in top right"
ggplot(dat, aes(x=factor(carb),y=horsepower,fill=factor(carb))) +
  geom_bar(stat="identity") +
  labs(title="", x="number of carburetors", y="mean horsepower")
```

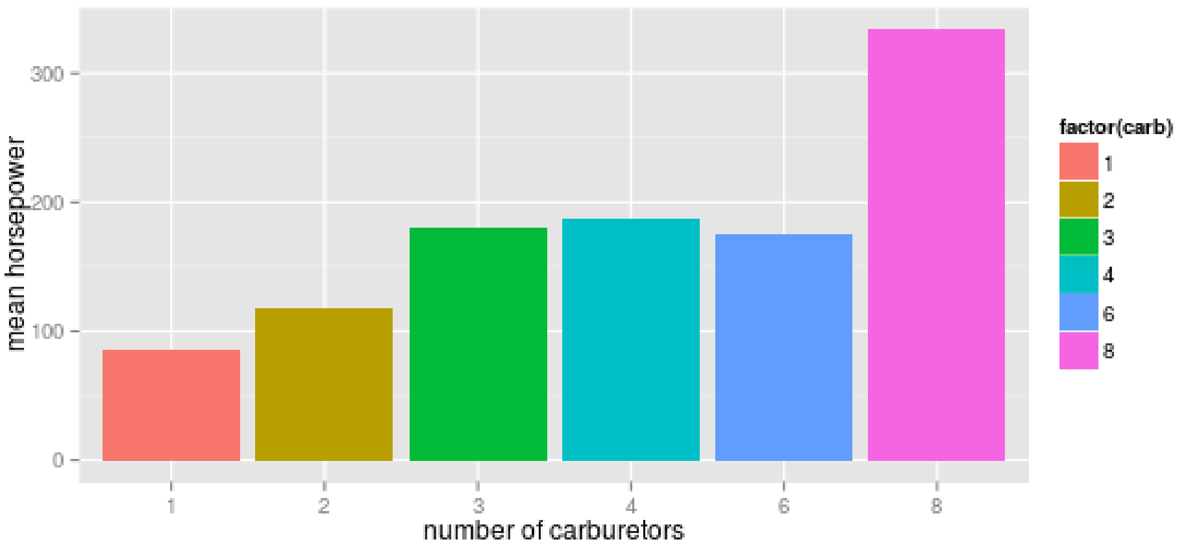

Figure 1: This is just on example plot

The Global Statistical Yearbook, which is the parent publication of the suite, provides a statistical synthesis of global and aggregate trends and an introduction to the major issues facing agriculture from the perspectives of food security, rural development, poverty, natural resource and environmental sustainability. The publication encompasses over thirty topic-based dimensions and employs over 350 indicators. Later this year, the third global publication will be launched.

In line with the Organization's emphasis on decentralization, a common set of regional-focused Statistical publications provide a synthesis to the parent publication. The emphasis on regional concerns sets these products apart from the global publication, specifically by offering text, maps, tables and graphs that are

[1] http://www.fao.org/economic/ess/ess-publications/ess-yearbook/en/#.UbHn8Nhj7To

	Model 1	Model 2
(Intercept)	1.08^{***}	6.53^{***}
	(0.07)	(0.48)
Petal.Width	2.23^{***}	
	(0.05)	
Sepal.Width		-0.22
		(0.16)
R^2	0.93	0.01
Adj. R^2	0.93	0.01
Num. obs.	150	150
RMSE	0.48	0.83

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: A table example: Statistical models

relevant to the region. These yearbooks empower the work of FAO in the field by equipping the Regional Offices with a dissemination product of high added value. In 2014, all five regions launched a Regional Yearbook.

A derivative of the parent publication, the Pocketbook, serves as a rapid and accessible reference point on global food and agricultural trends, supported by individual country profiles. A Pocketbook will be produced every other year to compliment the Yearbook. In 2014, FAO launched "Food and Nutrition in Numbers," which came out at the Second International Conference on Nutrition (ICN2). This year, FAO is expected to produce five regional Pocketbooks.

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[2] in combination with the typesetting program LaTeX[3]. Both are free and open source software5. The two programs work in concert to provide a seamless process in generating data publications and therefore more efficient dissemination tools. The work is documented and kept organised using free and open source revision control system Git and the collaborative development is hosted in Github[4].

1.2 Motivation of using an open-source programming language

The idea of using free and open source programming language is two-fold: the first involves the possibilities of free software in general and second the ability to program the whole process from the raw data to finalized publication. In short term, migration from proprietary software into free software cuts the licensing costs, but may slow down the process if people are new to the software. However, in the longer run, once the users are capable of modifying the software, they can contribute also to the content of the software and not just the yearbook, and the whole potential of free software can be utilized. Publishing the improved software as free software makes the process transparent and allows other researchers to understand and replicate the process, and to contribute to the process themselves.

The second reason concerns the programmatic approach in preparing the SYB. Advantages of programmatic approach involve automation, the text based sources, communication and quality control. The high level of automation significantly reduces manual operations, which are often prone to human errors. Also, it saves significant time when the outputs have to be updated for instance due to changes in underlying raw data. Moreover, the fact that a programming code is essentially plain text allows the use of revision/version control for easily tracking the changes, implementing new features without breaking the existing workflow and maintaining various versions of the software.
Free software are a result of ongoing collaborative effort by the user communities. Most building blocks of the SYB workflow are taken from the R and LaTeXcommunities, and the new implementations are given back to the community for feedback, criticism and inspiration. As all steps of data manipulation and creation of

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

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

[4] https://github.com/

charts and tables are recorded in the code, problems and issues can be communicated to wider community of users as reproducible examples. Such crowd sourcing is typical in free software communities and increases greatly the changes for solving the issues and, at the same time, provides inspiration for other users. The credibility this transparent approach gives to the process is crucial in this context and may lead to a greater harmonization in dissemination of similar statistical information.

1.3 Motivation of using R and LaTeX

R is often 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. The community of R-users is growing both in number of users and contributed software extension know as packages. Only the official CRAN[5] repository has currently 6520 packages, whereas Github hosts roughly 1,500 packages for R. In addition, several domain specific projects such Bioconductor[6], rOpenSci[7] or rOpenGov[8] provide support and visibility for packages within particular domain across many fields of computational sciences. The freedom to study how the specific functions work allows the user to adapt the algorithms to specific needs and therefore to continuously improve packages. And finally, the high connectivity to other languages and software makes R a flexible tool that responds to a variety of users' needs.

The major downside of using R is the initial learning curve. In addition, the speed is sometimes 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 apply primarily to those who use it on a daily basis.

In the SYB process, R is used for data processing, analysis and dissemination. More specifically, R automatically downloads variables, imports datasets, merges datasets with different standards, constructs new variables and computes aggregations. It also disseminates data through charts, tables, and maps, which are automatically translated into format suitable for LaTeX.

LaTeXis an open-source high-quality typesetting system that includes features designed for the production of technical and scientific documentation. It is based on TeX[9] typesetting program. As with R, LaTeXis open source and therefore has no licence costs or license restrictions. Moreover, LaTeXis a mark-up language and thus shares many advantages of programming languages. Compared to other word processing programs like MS Word or LibreOffice Writer, LaTeXaccommodates mathematical notation easily, controls sectioning, cross-references, tables and figures, and automatically generates bibliographies and indexes. LaTeX's fundamental idea is to separate the content and the formatting, that allows the user to write the document without having to simultaneously control the formatting as with word processors. It allows multi-lingual typesetting and guarantees perfect consistency throughout a document/book (e.g. in the management of the colours and layout of the spreads), thus facilitating enormously the printing step without intervention from graphic designers. The challenge, however, is that there is also a learning curve, even though all the most popular text editors do support LaTeXmark-up and there are even specific software for it such as Lyx[10].

LaTeXis used to typeset the entire SYB publication. The dissemination objects translated into LaTeXcode by R are automatically included and formatted within the publication through the specific class `faoyearbook`. This package defines all of the needed commands to delineate the structure and build the publication. The package is geared towards the Yearbook but can be adapted 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. The most common technology is Git[11], but there are several other solution available for developers. Revision control system allows users to revert to previous versions of a document and to track all the modifications within the file. In addition, users can work

[5] http://cran.r-project.org/web/packages/

[6] http://www.bioconductor.org/

[7] https://ropensci.org/

[8] http://ropengov.github.io/

[9] http://tug.org/

[10] http://www.lyx.org/

[11] http://git-scm.com/

simultaneously on the same project on multiple computers, even offline, as each computer has it's own copy of the source code. When using revision control the developers maintain a stable and working master version of the software and create their own copies of the master called branches for developing new features. Once tested compatible with the master of the project, the branches can be merged into the master adding the new features. A complete history of what was added or removed by whom remains in the version history and the project can be reverted to any point of time if needed. R and LaTeXscripts are essentially text files that are ideal for revision control unlike binary files such as Excel sheets or pdf documents. In the SYB process, , Git is used for revision control and GitHub for hosting the collaborative work on R-packages and LaTeX-class . Actual project folders and raw data is synced between multiple computers using Bittorrent Sync[12].

2 SYB production

The construction of the SYB revolves mainly around processing, analysis and dissemination steps, which are described by the Generic Statistical Business Process Modell[13] and illustrated in figure 1. The process is divided in three sub process of which begin with the construction of database for the SYB-production (FAOSYB-database). Once database is constructed takes place the creation of the dissemination object, ie. graphs, maps and tables. Once ready the dissemination object are complemented with interpretive text and vowed into publication ready for printing and electronic dissemination. Database construction and creation of the dissemination object are processed entirely with R, while the publication is produced with LaTeX. More specifically, two R packages have been created. The FAOSTAT package hosts a list of functions to download, manipulate, construct and aggregate statistical data, while the FAOSYB package includes functions for creation of the dissemination objects.

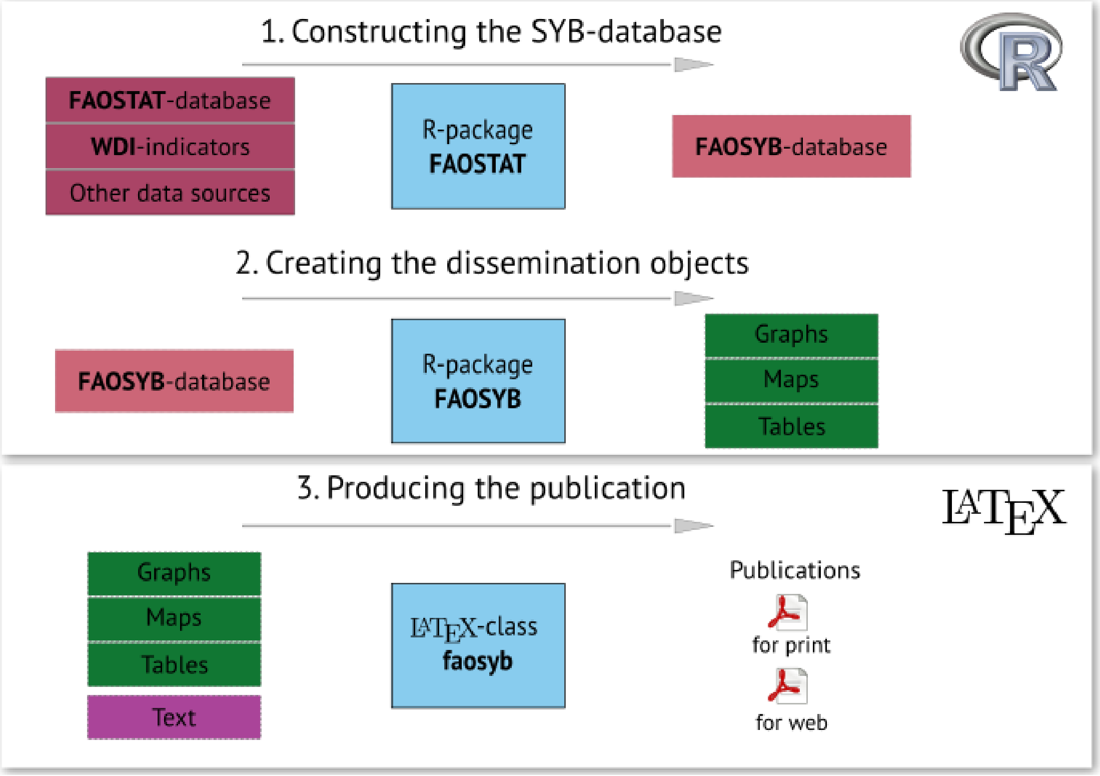

Figure 2: Statistical Yearbook process

[12] https://www.getsync.com/

[13] Following the GNU definition, "'free software' means software that respects users' freedom and community. Roughly, the user has 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" (https://www.gnu.org/philosophy/free-sw.en.html)

2.1 Constructing the SYB-database

2.1.1 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 must manually import the data. The FAOSTAT package provides the functions to automatically download variables from the FAOSTAT[14] and the World Data Bank (World Development Indicators)[15] databases as well as the algorithms needed to manually import specific sets of data. Although not yet implemented in this process, it is also possible to retrieve data through SDMX with R.

2.1.1.1 Data retrieval 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` provides access to FAOSTAT data through the FAOSTAT API. The user is facilitated in the construction of the API query by the function `FAOsearch`, which, with very few steps, gives users needed codes to build the query. Similarly `getWDI` and `getWDItoSYB` are functions to download the World Development Indicators.

2.1.1.2 Data retrieval from manual sources Non-harmonised manually constructed data comes in many structures and file formats and can be challenging to harmonise with the data retrieved from databases. Greatest concern, however, arises from how the countries are classified in the data, as different organisation do not share a common classification system. Therefore all the (frequent) exceptions to the reference country classification system have to be taken into account in order to accurately design a software to address these issues.

There are four main issues in attempting to combine different country classification systems. The first is differences in the country definitions, which are generally due to varying legal entities. The second is changes in the country composition over time of which China is one example. The World Bank disseminates China (excluding Taiwan), Hong Kong and Macao, whereas, until recently, FAOSTAT disseminated only China (including Taiwan, Hong Kong, and Macao). It is clear that, in the two cases, "China" represents different realities. Third 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 during the same year can also include data for South Sudan, therefore leading to double counting. Fourth, the analysis is further complicated by disputed territories and economic unions, such as Ilemi Triangle, which do not have representation under most of the country coding systems.

A precise matching is thus essential. The SYB uses the M49 country classification system[16] following the idea of a needed convergence to a common international standard. 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 matching is not always possible.

2.1.2 Merging and scaling the data

2.1.2.1 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.

2.1.2.2 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

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

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

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

and results to be obtained. This heterogeneity would imply a complex matching of different systems that, for the moment, has not been developed.

2.1.3 Aggregating the data

Aggregation is another data manipulation step that is commonly overlooked, especially when data are taken from different sources. In most of cases the already computed aggregates cannot be used because a) a different country classification system is used; and/or 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. In the Yearbook production, such aggregates are not discarded but are used to check the aggregates computed for potential errors in the methodology and/or weighting variables.

To overcome these issues the data must be first converged with a specific country classification system, and thereafter new consistent and comparable aggregates must be generated. In SYB process this issue is addressed using a two-step aggregation process at 1) country level and 2) geographic and economic level.

2.1.3.1 Aggregation at country level The starting point in the aggregation process is a set of countries that is as disaggregated as possible through the FAOSTAT coding system. The objective is then to match the information with the M49 country level definition. Therefore, in the first aggregation step countries that go together following the M49 system and merged together, such as Tanzania and Zanzibar. The `aggCountry` function aggregates territory entries into countries or higher level classifications based on the relationship specified.

2.1.3.2 Aggregation at geographic and economic level In order to compute geographical, economic, and political aggregates, a further aggregation step is necessary. The first problem is that a hierarchical approach cannot be used, though it would be a logical step to compute the aggregations starting from the set of countries obtained after the first aggregation.

First issue is that 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. One such example is Reunion, which is part of France yet is located close to Africa. The second issue is that, following the hierarchical approach would risk excluding the time dimension, and therefore 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, so there are no challenges for a current year. However, given that we are interested in computing aggregates for past years, we need to consider how the world composition has evolved 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 would not be valid in 2010, when South Sudan and the Sudan were a single country. For this reason, the aggregate for Africa can only be computed consistently by including South Sudan, Sudan and Sudan (former). The presence of data for all the three Sudans would imply a double counting. Therefore the hierarchical approach cannot be applied.

However, geographic and economic level aggregation step also begins from the large set of countries, but needs a partially different relationship. Two rules are implemented to ensure that 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 (therefore the country composition can change during a specified period by up to 15 countries) because it does not make sense to compare aggregates for two different years if the number of reporting countries vastly differ. Both of these rules are automatically applied by the function `aggRegion`.

2.1.4 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-à-vis the total. There are also two types of growth rates within the package: the least squares and the geometric growth rates. 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 requires a base year, while the second the year interval.

2.1.5 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 on to the user through a specific object, sub-section and section. It should 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 has the risk of not having enough data to calculate many aggregates. 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 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.

2.2 Creating the dissemination objects

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 colors 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[17]. Tables are generated using internal codes and have not yet been added to the package due to their complexity. However, what is important is that the generation of charts, maps, tables and mini-tables are essentially implemented as R code, and, for this reason are easily reproducible and updatable. In the end, objects, captions, sources and metadata are automatically translated into LaTeXformatting.

2.3 Producing the publication

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`. LaTeXcontrols automatically the sectioning, cross-references, and indexes. The bibliography is done with BibTeX[18] and it is read automatically by LaTeX.

3 Conclusion

Up until now, many statistical yearbooks have been produced with a fairly large team that manually downloads data and inputs the information. Such workflow, however, is prone to error, labour intensive and often expensive. FAO, through the production of its Statistical Yearbook, offers an alternative. Two open source software, R and LaTeX, are used in all steps of the process: retrieving and merging data, conducting computations, creating visualizations and managing the layout. All steps are documented with code and are therefore reproducible and transparent, and can be used by anybody in the production of other publications. The LaTeXclass faoyearbook defines all of the needed commands to delineate the structure and build the publication.

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

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

4 References

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