A Pcakge for Transparent and Reproducible Statistics: Package FAOSTAT

Michael. C. J. Kao
Food and Agricultural Organiazation
of the United Nations

Abstract

The aim of this document is to introduce the FAOSTAT package developed by the Food and Agricultural Organization of the United Nations which serves as the core back bone for the production of the statistcal year book.

Dealing with official statistics is a very tedious and complex task which is usually overlooked. This paper will address some of the common problems we have encountered and at the same time demonstrate how some of these problems can be alleviated by the package and provide a framework for reproducible statistics.

The use of open source software R and LaTeXbrings tremendous amount of benefits, speeding up the production process and open up the data and methodology to the general public.

In this small document we will illustrate the production process and demonstrate how the use of the package can increase transparency and sustainability. Furthermore, we will point out some details which are typically overseen by analysts and researchers.

Keywords: R, Official Statistics.

1. Introduction

The idea of using R and LaTeXfor the production of the FAO statistical year book was initiated by Adam Prakash and Mattieu Stigler in the ESS division of the Food and Agricultural Organization of the United Nations. The initiative was taken in order to replace the labour intensitive work with a streamline system which integrates data extraction, maniapulation, statistical graphics and tables into one single integrated system.

This paper demonstrate how to use the FAOSTAT package to download, and process data under the framework of the FAO statistical year book. This will serve as a template in which analysts and researchers can used and modify to suit their needs.

First, we will demonstration the useage of the `getWDItoSYB` and `getFAOtoSYB` functions to download data from the FAO FAOSTAT and the World Bank WDI API. We then continue by how the `mergeSYB` function can be used to combine data from various sources. This is then followed by method available to compile new statistics from the source data and finally privide an example of how to aggregate data.

2. Motivation

Compiling hundreds of statistics from different sources under traditional approach such as Excel can be very labour intensive and error proned. Furthermore, the knowledge and the experience is almost impossible to sustain in the long run resulting in inconsistent result and treatment over time. As a result, the ESS took the initiative to use R and LaTeX as the

new architecture for a sustainable and cost-effective way to produce the statistical year book. This approach increases the sustainability and coherence of the publication as all the data manipulation, and treatment are recorded in the source code.

In addition to these working motives, the use of R enables the data generated by the publication to be reproducible and readily accesible to researchers and analysts around the world. This open-data philosphy has proven to create tremendous amount of benefits for both the user and the data provider. We hope that this initiative will increase the visibility of aricultural related statistics and spark more research and analysis which the organization and its beneficiaries will gain.

After a little more than a year, the architecture has not only become the standard for producing the global statistical year book and the pocket book; but the same code is used as template to generate five other regional books. The gain in efficiency was unimaginable with three individual is sufficient to generate four publication including the text within a year and potentially raised to seven this year.

Reproducibility is the norm in the academic. This property allows one to verify, improve and reproduce the research for future use.

We believe that a publication such as the statistical year book which publishes statistics and aggregates should be examined under the same transparency standard. It is important not only to publish these figure, but also at the same time the methodology in which the data are prepared. Furthermore, we hope the publication of the software along with the experience will gain from valuable feedback which will serve as potential ground for a harmonized framework.

```
## if(inherits(try(find.package("FAOSTAT")), "try-error"))
##    install.packages("FAOSTAT")
library(FAOSTAT)
```


3. Download data from FAOSTAT

FAOSTAT is the largest agricultural database, it contains data from land productivity to agricultural production and trade. More information can be found on the official website of FAOSTAT http://faostat3.fao.org/home/index.html and the Statistics Division (ESS) http://www.fao.org/economic/ess/en/

The Domain, element and item code is required to download the data, an interactive function has been provided to the user to find the respective codes. After the search, the codes will be stored as `.LastSeach` to streamline the download.

```
FAOsearch()
test = getFAO(query = .LastSearch)
```

The `getFAOtoSYB` is a wrapper for the `getFAO` to batch download the data. It supports error recovery, and stores the status of the download. The function also splits the data downloaded into country level and regional aggregates, saving time for the user.

```
FAOmetaData.df = data.frame(varName = c("arableLand", "cerealExp", "cerealProd"),
                            domainCode = c("RL", "TP", "QC"),
                            itemCode = c(6621, 1944, 1717),
                            elementCode = c(5110, 5922, 5510),
                            stringsAsFactors = FALSE)
```

```
FAO.lst = with(FAOmetaData.df,
    getFAOtoSYB(name = varName, domainCode = domainCode,
                itemCode = itemCode, elementCode = elementCode))
```

The object returned is a list of length three, these are entity level data, aggregates and the result of whether the download was successful.

4. Download data from World Bank

The World Bank also provide an API in which provide a vast amount of data open to the public. More information about the data and the API can be found on the official website http://data.worldbank.org/

The author is aware of the `WDI` package available on CRAN, however it was developed under the use for World Bank while we have tailored it so that it can be used along with other sources of data and also for large batch download.

```
WB.lst = getWDItoSYB(indicator = c("SP.POP.TOTL", "NY.GDP.MKTP.CD"),
                     name = c("totalPopulation", "GDPUSD"),
                     getMetaData = TRUE, printMetaData = TRUE)
```

5. Merge data from different sources

Merge is a typical data manipulation step in daily work yet a non-trivial exercise especially when working with different data sources. The built in *mergeSYB* function enables one to merge data from different sources as long as the country coding system is identified. Currently the following country coding translation are supported and included in the internal data set FAOcountryProfile of the package:

- United Nations M49 country standard [UN_CODE]
  http://unstats.un.org/unsd/methods/m49/m49.htm
- FAO country code scheme [FAOST_CODE]
  http://termportal.fao.org/faonocs/appl/
- FAO Global Administrative Unit Layers (GAUL).[ADM0_CODE]
- ISO 3166-1 alpha-2 [ISO2_CODE]
  http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2
- ISO 3166-1 alpha-2 (World Bank) [ISO2_WB_CODE]
  http://data.worldbank.org/node/18
- ISO 3166-1 alpha-3 [ISO3_CODE]
  http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3
- ISO 3166-1 alpha-3 (World Bank) [ISO3_WB_CODE]
  http://data.worldbank.org/node/18

Data from any source with the classification listed above can be supplied to `mergeSYB` in order to obtain a single merged data. However, the column name of the country coding scheme is

required to be the same as the name in square bracket, the responsibility of identifying the coding system lies with the user.

Below we list some commonly observed problem when merging data from different sources.

5.1. Identification problem

Due to the fact that different organization are bounded by different political agenda the user need to be aware of the precise definition of the country and also the legal recognition.

For example, the China provided by the World Bank does not include Taiwan, Hong Kong and Macau. On the other hand, FAO provides not only a single China (FAO = 41), but also China plus Taiwan (FAO = 357) depending on the context. In addition, it is common to observed statistics for China (ISO2 = CN or ISO3 = CHN) which includes Taiwan, Hong Kong and Macao. The default translation matches China to China mainland (FAO = 41).

5.2. Representation problem

Moreover, the situation is further complicated by disputed territory or economic union such as Kosovo and Belgium-Luxembourg which does not have representation under certain country coding system.

```
FAO.df = translateCountryCode(data = FAO.lst$entity, from = "FAOST_CODE",
    to = "ISO2_WB_CODE")

## Warning:  Please make sure that the country are matched according to their definition
## Warning:  The following entries does not have 'ISO2_WB_CODE' available

##       FAOST_CODE ISO2_WB_CODE
## 680           15         <NA>
## 2618          62         <NA>
## 10295        351         <NA>
## 10346        357         <NA>
##                                              OFFICIAL_FAO_NAME
## 680                                         Belgium-Luxembourg
## 2618                                              Ethiopia PDR
## 10295 China (China mainland, Hong Kong SAR, Macao SAR, Taiwan)
## 10346                          China (China mainland, Taiwan)


WB.df = translateCountryCode(data = WB.lst$entity, from = "ISO2_WB_CODE",
    to = "UN_CODE")

## Warning:  Please make sure that the country are matched according to their definition
## Warning:  The following entries does not have 'UN_CODE' available

##      ISO2_WB_CODE UN_CODE OFFICIAL_FAO_NAME
## 5513           KV      NA            Kosovo
```

5.3. Transition problem

Finally, the discontinuity and transition of countries further increases the complexity of the data. The South Sudan was recognised by the United Nations on the 9th of July 2011, however,

the statistic reported by the Sudan in the same year can also includes data for South Sudan and thus failing the mutually exclusive test. Further more if the aggregates are computed for periods prior to 1992, then a region definition which does not back track in time will not include Former Yugoslavia and thus the aggrgates computed will be incorrect. For more details about historical and transitional countries please refer to http://unstats.un.org/unsd/methods/m49/m49chang.htm

Given the lack of an internationally recognised standard which incorporates all these propoerties, we suggests the use of the FAO country standard and region profile shipped with the package which addresses most of these problems.

```
merged.df = mergeSYB(FAO.lst$entity, WB.lst$entity, outCode = "FAOST_CODE")

## Warning:  Please make sure that the country are matched according to their definition
## Warning:  Please make sure that the country are matched according to their definition
```

6. Computing growth, and other derivatives

There are two types of growth rate shipped with the package, the least squares growth rate and the geometric growth rate. The least sqaures growth rate is used when the time series is of sufficient length. The default is at least 5 useable observaions, however if the time series is sparse and more than 50% of the data are missing than the robust regression is used.

```
newData.df = constructSYB(merged.df, origVar1 = "totalPopulation",
                          origVar2 = "totalPopulation",
                          newVarName = "totalPopulationGrowth",
                          constructType = "growth",
                          grType = "ls", grFreq = 10)
```

7. Aggregation

Aggregation is another data manipulation step that is commonly over seen. The result can vary due to the difference between the regional definition and the set of countries used. Furthermore, it is complicated by the amount of missing values which can render the aggregates incomparable. Given the missing values and country sets, aggregation can only serve as approximates in order to inform the general situation of the region. The following rules are implemented to ensure the aggregates computed are meaningful and comparable.

- A minimum threshold in which the data must be present, the default is 65%.
- The number of reporting entities must be similar over the years. It does not make sense to compare aggregates of 1995 and 2000 if the number of reporting countries differ vastly, the default tolerance is 15.

In addition, historical countries are aggregated to ensure comparability over time. For example, The Former Soviet Union is not part of the current definition of the M49 standardard, nevertheless, it would be ignorant to omit it from the aggregation.

```
relation.df = FAOregionProfile[, c("FAOST_CODE", "UNSD_MACRO_REG_CODE")]

region.df = aggRegion(data = merged.df, relationDF = relation.df,
                      aggVar = c("arableLand", "cerealExp", "cerealProd",
                                 "totalPopulation", "GDPUSD"),
                      aggMethod = rep("sum", 5))
```

Affiliation:

Michael C.J. Kao
Economics and Social Statistics Division
Economic and Social Development Department
United Nations Food and Agriculture Organization
Viale delle Terme di Caracalla 00153 Rome, Italy
E-mail: michael.kao@fao.org