faoswsFlag:A package to manage flag aggregation

Michael. C. J. Kao
Food and Agriculture Organization
of the United Nations

Abstract

This short documentation is intended to explain how observation flags are aggregated in the ESS Statistical Working System.

Keywords: meta data, flag aggregation.

Lets start by loading the required library into R.

```
## Load the required libraries
library(faoswsFlag)
library(ggplot2)
```

Since the introduction of the new statistical working system, the old symb which reprents how the data was collected and computed is split into two separate flags. The first, an observation flag which is a description of the observation status, whether it be official, estimates or imputed value. While on the other hand, the methodology flag contains information of how it is collected or computed. It can be from survey, questionnaire or it can be obtained as a balance or estimated through statistical methodology.

This paper concerns mainly the observation flag, where the data originated and the information content it is associated with.

Shown below is the corresponding table for the observation flags as of September 1, 2014 outlined in Annex 5 of FAO Statistical Standards:

Table 1: Description of the Observation flag

Flags	Description
<blank>	Official Figure
E	Estimates
I	Imputed
M	Missing
T	Unofficial figure

One of the main goal for this package is to provide a consistent framework for handling multiple flags and flag aggregation.

In order to compute aggregation of flags, one must convert the symbol into a numerical type. The way how this is handled is based on the amount of information quantity and the reliability of the observation status.

Information obtained from reliable source should have a high information content and thus should be converted to a high value, while data estimated or derived should have a lower information vice versa.

Shown below is the default weights table for the new statistical working system which comes along with the package.

```
## Printed here is the default flag conversion table shipped with the
## package.
faoswsFlagTable

##   flagObservationStatus flagObservationWeights
## 1                                         1.00
## 2                     T                   0.80
## 3                     E                   0.75
## 4                     I                   0.50
## 5                     M                   0.00
```

From this table, we have assigned 1 to official figure while 0 to missing values. Albeit the arbitrary selection of the values, it provides a rank of the information content which is what we need in order to compute flag aggregation.

Take the computation of yield for example, the value is computed based on production and area harvested which may come from different sources. Thus, when we compute a derived statistic which is unobserved such as yield it is important to reflecte the information content which associated with this derivative. For a set of aggregation, the minimum of the set is taken as the final observation flag.

Lets say we have a production value of 30,000 recorded from official survey (), while the area harvested was collected from an unofficial private data base (T). As shown in the following code, the resulting flag for yield should return (T) to reflect the lower information content of the unofficial figure.

```
## The function works just like sum(), with an optional arguement for
## the flag table to be used.
aggregateObservationFlag("", "T", flagTable = faoswsFlagTable)

## [1] "T"
```

```
## Aggregation of multiple flag

## Simulate flag for production
simulatedProductionFlag =
    faoswsFlagTable[sample(1:NROW(faoswsFlagTable), 10, replace = TRUE),
                    "flagObservationStatus"]
simulatedProductionFlag

##  [1] "T" "E" "E" "T" "I" "I" ""  "E" "I" "I"

## Simulated flag for area harvested
simulatedAreaFlag =
    faoswsFlagTable[sample(1:NROW(faoswsFlagTable), 10, replace = TRUE),
                    "flagObservationStatus"]
simulatedAreaFlag

##  [1] "E" "E" "I" "M" ""  ""  ""  "M" ""  "M"
```

```
## Now compute the aggregation of flag
aggregateObservationFlag(simulatedProductionFlag, simulatedAreaFlag,
                         flagTable = faoswsFlagTable)

##  [1] "E" "E" "I" "M" "I" "I" ""  "M" "I" "M"
```

Currently, the flags are chosen as arbitrary mainly to preserve a rank order based on expert judgement. Nevertheless, this information can be estimated from the data and history of the flag. This is currently under investigation, yet the idea is simple. First, we take official figure as the golden benchmark, then we calculate the deviation of each observation which has a value collected priorly with a different flag. This will allow us to compute the relative error of flags and in turn to estimate the information content.

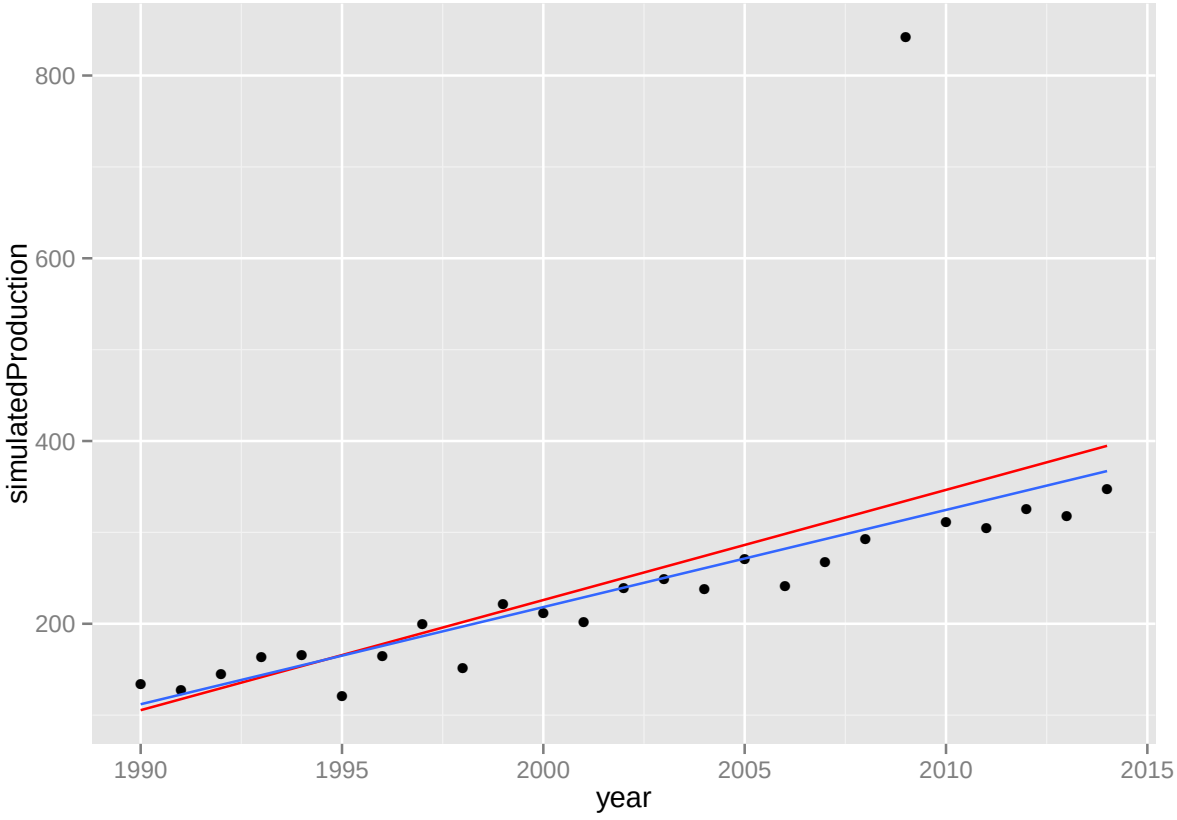
Associating a value with its meta data has significant advantages, as it provides information about the data which can be used in further analysis or improvements of modelling. For example, instead of fitting a linear regression by treating all observation equally, we can estimate a weighted regression which gives more weight to data which are of higher reliability.

The following artificial example illustrates how accounting for the information source can result in a better fit and incorporate poor data quality. The artificial data starts in 1990 and ends in 2014, with all the observation collected as official figure except for 2009. For illustrative purpose, the value in 2009 was imputed by a poor algorithm and can be seen in the graph as an outlier. The illustration shows how accounting for the outlier through the use of meta data can result in more robust model fitting than as treating all data have the same information quality.

The black dots are the original data, with the red line being the linear regression treating all observation equally while the blue line incorporates the information quality through the flags. We can see that by accounting for the source, the fit of the blue line is less affected by the outlier generated by the poor imputation algorithm.

```
## Simuate a data set which has a single point that was imputed badly
## but still used for later analysis.
x = 1990:2014
y = 100 + 10 * (x - 1989) + rnorm(length(x), sd = 20)
f = rep("", length(x))
y[20] = y[20] + 500
f[20] = "I"
simulated.df = data.frame(year = x, simulatedProduction = y, flag = f)

## Plot the data and show the two different fit when accounting for the
## source and quality of information.
ggplot(data = simulated.df,
       aes(x = year, y = simulatedProduction)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x,
                data = simulated.df, se = FALSE, col = "red") +
    geom_smooth(method = "lm", formula = y ~ x,
                aes(weight = flag2weight(flag)),
                data = simulated.df, se = FALSE)
```



Affiliation:

Michael. C. J. Kao
Economics and Social Statistics Division (ESS)
Economic and Social Development Department (ES)
Food and Agriculture Organization of the United Nations (FAO)
Viale delle Terme di Caracalla 00153 Rome, Italy
E-mail: michael.kao@fao.org
URL: https://github.com/mkao006/sws_imputation