Flag Aggregation and Much More ...

Michael C. J. Kao

Food and Agriculture Organization
of the United Nations

Outline

Outline for section 1

Background

Since the introduction of the new Statistical Working System (SWS), many innovative approaches has been devised to improve the current status and to accomodate for future needs of the fast changing world of statistics.

One subtle yet fundamental change is the separation of a single symbol into two separate flag. A symbol or flag is a meta data which indicate the collection or methodological procedure that generates the value.

The aim of this presentation is to introduce the change and how the observation status flag can impact future works.

Mixing of Status and Method

Historically, a symbol can represent how it was calculated, or where it was is collected. However, this mixed approach has created some confusion and loss of information.

For example, when yield are calculated based on production and area harvested, a flag "C" representing "calculated" is assigned. However, it does not show the observation status of yield.

A value for yield calculated based on official data has the same equivalent meaning to those calculated on estimates.

This mixing of information results in loss of information and potential bias analysis.

Separation of Status and Method

To better represent these information, the decision was made to split the symbol into two separate flag reflecting different piece of information.

One for the **observation status** which represents how the data was observed, whether collected from official or semi-official data source, or it maybe estimated or imputed.

The second flag would denote the **methodology** it was obtained. Official figure can be obtained from questionnaire, or it can be from database or publications. Value which are estimated can be manually derived based or algorithm driven.

Problem

Nevertheless, it presents a problem. How do we assign observation flag when a value is calculated? Take the yield for example again, we would assign a method flag indicating the value was calculated; but what is its observation status?

Outline for section 2

What should the correct observation flag?

One approach is to quantify the believability of information for each flag and treating them as ordinal variables. Then the aggregation can be proceeded by taking the lower bounds of the set.

The rational for that is that the believability of an aggregation should reflect the lowest level of believability.

The Flag Table

Shown below are the weights used to rank the flags.

Table : Description of the Observation flag

Flags	Weights	Description
(blank)	1	Official Figure
T	0.8	Unofficial figure
E	0.75	Estimates
I	0.5	Imputed
M	0	Missing

An example

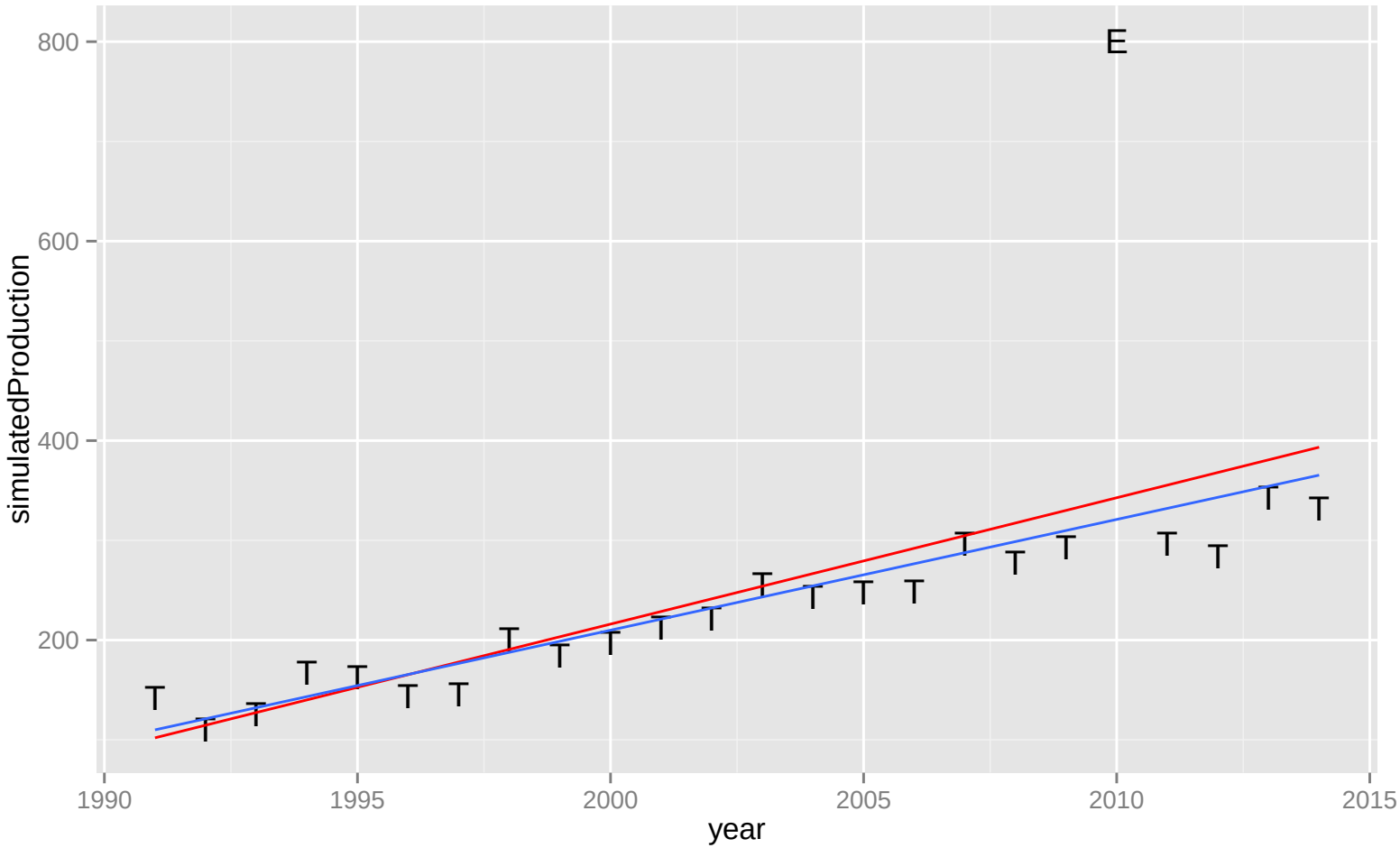
Given that the yield is computed based on production and area harvested, then the observation flag should only depend on the flag of production and area harvested and is the minimum of the two.

Lets assume that official figure are of higher quality than unofficial sources. Then the flag of yield computed from a production collected from official survey and area harvested collected from unofficial database should be unofficial.

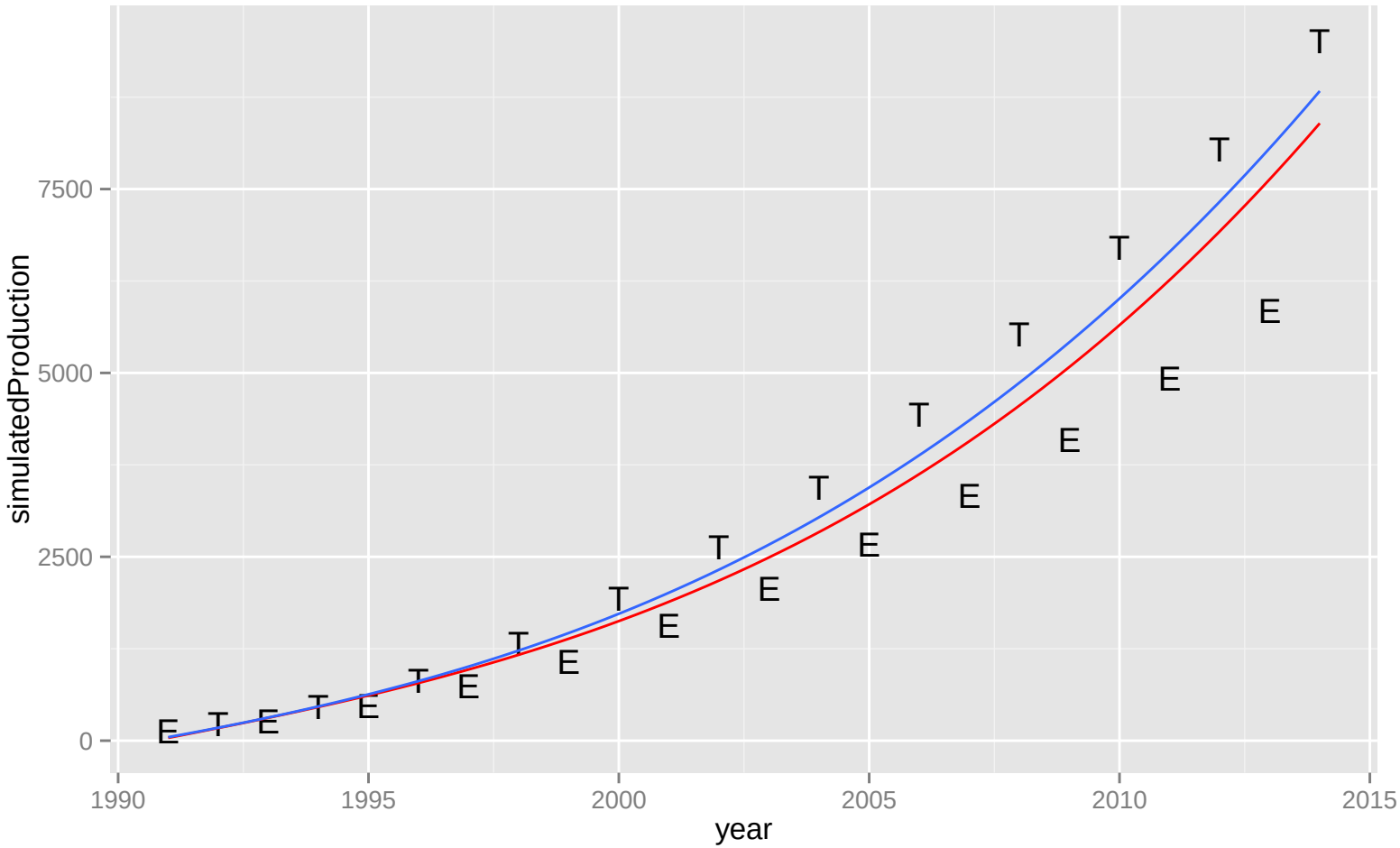
Outline for section 3

The recognition of various data source and quantification of the believability provides many potential application which can improve the quality of overall statistics.

Robust From Outlier



Combining Source of Different Information Quality



The potential associated with quantifying the quality of information is huge. It provides an additional piece of information for every single data point.

This is a general framework, no specific methodology is required since any model and methodology can utilize this framework by incorporate the additional information.

Outline for section 4

Currently, the weights are assigned subjectively by expert judgements in order to preserve a rank order.

Table : Description of the Observation flag

Flags	Weights	Description
(blank)	1	Official Figure
T	0.8	Unofficial figure
E	0.75	Estimates
I	0.5	Imputed
M	0	Missing

However, methods are being devised to estimate these weights objectively and automatically.

A requirement for automatized and objective computation of weights is required due to the fact that we do not have the human resources to devote to assigning weights for each country, each agricultural domain and partition of the data.

Below we present two methods that are under-investigation for discussion.

Measuring Loss of Information

In this method, we take the official figures as gold standard. Then we compute the Kullback-Leibner Divergence and measure the amount of information loss when we represnt value in alternative sources other than official source.

Kullback Leibler Divergence

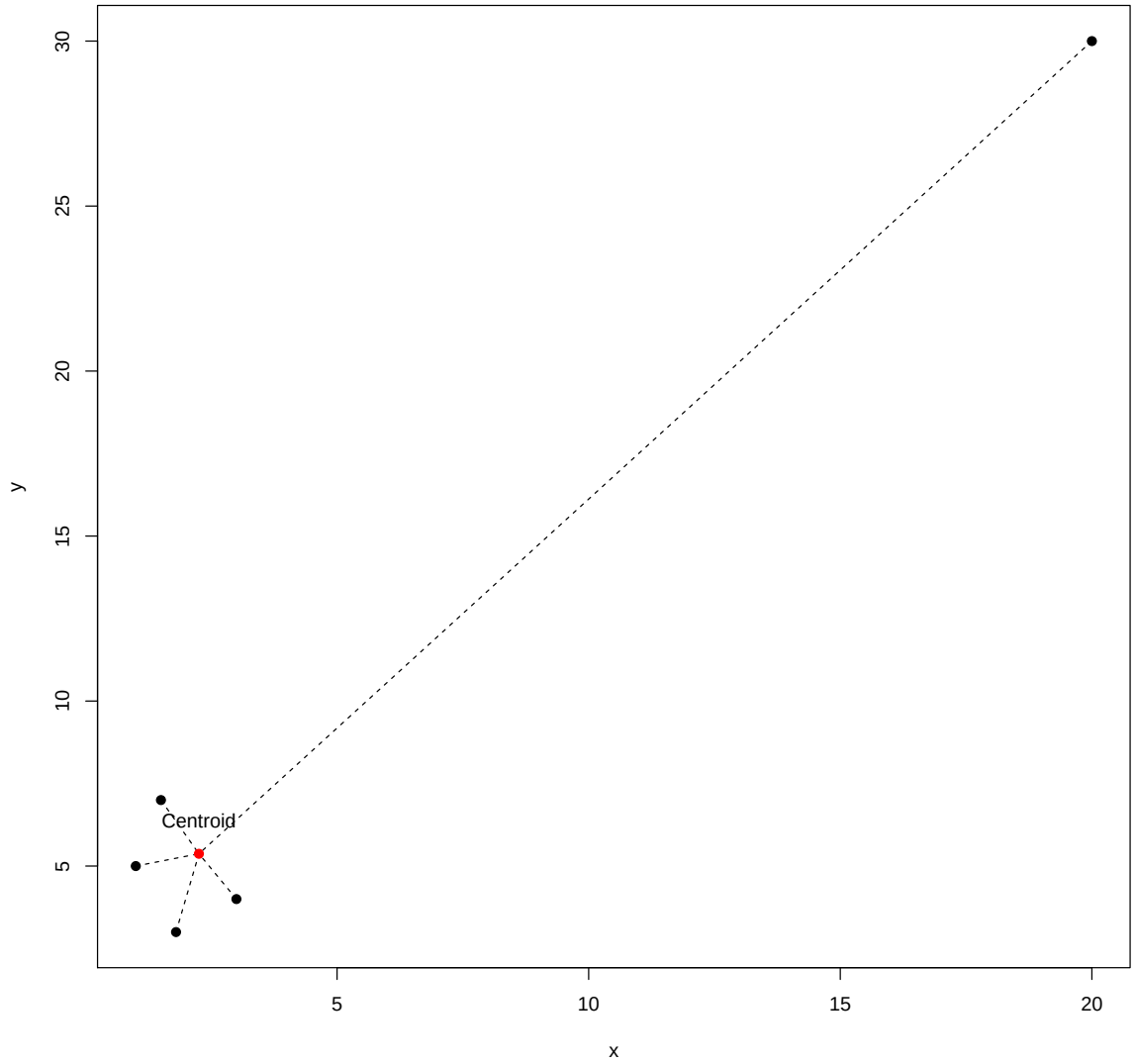
$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i).$$

The weights can then be calculated based on the relative loss of information. The greater amount of information lost, the lower the weight.

Self-Similarity Weights

This alternative method does not assume which source is correct, rather it uses similarity between the observations of different flag to identify a centroid. This centroid is interpreted as close to the true value.

The distance from the centroid will be inversely related to the weight. That is, the further away you are from the centroid, the lower the weight will be assigned.


Centroid
y
x
5
10
15
20
5
10
15
20
25
30