Measuring and Minimising Information Loss in Source Aggregation

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

Abstract

Keywords: meta data, flag aggregation.

1. Introduction

International organizations such as FAO inevitably collects information and data from various sources and channels. In order to provide a comprehensive status of the world in their respective domain of operation, one must utilize as much information as possible.

Since the introduction of the statistical working system in the 1990's, flags representing different source of information has been recorded. This practice enabled the clerk to identify whether a value complies with the organization policy of official values are the most trustworthy, and also based on experience which source is more reliable. It also underlies some of the earlier algorithm implemented in the system whether a value should be calculated or remain as it is.

Yet, no attempt has been made to harmonize and treaet the unequal information content of various source. Data collected from different source and channel does not have the same information quality. Further, the various nature of the collection may give arise to different statistical properties. The user should not assume that the data are of equal quality as they may and will undermine the validity of the analysis.

The proposed methodology makes the first pursuit to provide a systematic approach for this problem. The aim of the approach is to allow information from separate source to represent a harmonized picture when combined to provide the status of the world.

This paper comprises of three sections. First, the motivation and the problem at hand is presented and why such a framework is necessary. Secondly, we present the theory where the fundamentals of the principle of minimum discrimination information is introduced. Then we illustrate the use of this framework and demonstrate how estimators are superior when information quality are addressed. Finally, a short conclusion and further work concludes the paper.

2. Motivation and Problem Statement

One of the problem first faced was to identify over the spectrum of the data collection method, which is more reliable and should be disseminated. The identification will enable us to decide which statistic to disseminate allowing our user to access the most reliable set of information.

Often an arbitrary approach is taken, some times based on experience or perception. For example, manual estimation is current perceived as better than algorithmic imputation. However, this may notnecessarily be true in particularly seeing the advacement of imputation methods which has proven to out perform human estimation. A formal assessment framework can assist us in identifying the better source for dissemination.

Further, when data are collected and disseminated, information were often treated as equal and assume to behave identically. Despite being the commonly acceptable practice, it is far from desirable. Data collected from various source and channel can behave vastly different depending on the tools employed. Land survery in contrast to estimation based on satelite images can produce very different figure for the same object of measurement.

Likewise everything we attempt to meausre, there is an uncertainty in the measurement under the influence of noises. Whether it be passage of time or the circumference of the world in the past, there are some measurement uncertainty.

Various level of technological development lead to different data collection practice. It would be infeasible to devise a standard for which all data collection are equivalent, thus a framework to enable users to account for data of various information quality is indispensible.

Another common problem encountered when derived statistics were computed, a common task in publishing official statistics. Vast amount of data are collected by agencies and international organization. To effectively present the data for policy formulation and monitoring to high level executives, aggregation and summarization is essential.

In the current working system, an aggregation is associated with the flag of "C" representing that the value was computed. However, this results in a large amount of information loss. Information on the sources of data collection is unpreserved and lost.

To resolve this problem, an entropy approach was taken to quantify information in order to measure the information loss and further minimize the information loss in the process of aggregation of information.

3. Theory

Information theory has its root in communication, where Claude E. Shannon laid the foundation of the field with the paper "The Mathematical Theory of Communication".

The quantification of information enables us to determine which source to disseminate as we are restricted to disseminate only one measurable value.

In order to quantify the information, we assume data collected by official sources are correct and perceive it as the signal under perfect condition. This is in concordance with the organization's policy where official sources are taken as golden measure. Based on this assumption, we can proceed with calculating the information loss when data from a different source is used in the absence of official data.

3.1. Principle of Minimum Discriminant Information and Cross-Entropy

Given derived information set, a new distribution q should be chosen which is as hard to discriminate from the original distribution p as possible; so that the new data produces as small an information gain as possible.

In another word, the principle states that if we have to choose another representation when official figures are unavailable, the information set which result in the least amount of information gain or uncertainty should be chosen.

Provided this, we can choose between which information is under less influence of noise and better represents the true quanity of measurement.

$$\mathrm{H}(P,Q) = \mathrm{H}(P) + D_{\mathrm{KL}}(P\|Q)$$

Where

$$H(P) = -\sum_i p(x_i)\log p(x_i),$$

and,

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

The quantity $\mathrm{H}(P,Q)$ is known as cross-entropy and it is the difference between the entropy of the true distribuion $H(p)$ and the information lost measured by the Kullbak-Leibner $D_{\mathrm{KL}}(P\|Q)$.

In practical cases, the empirical Kullback-Leibner is calculated. Values are binned and the relative frequency is used to approximate the real density $p(x)$ and $q(x)$. That is, the index i is not a single point, rather it is a binned density.

This quantity inform us about the information loss associating with the employ of an alternative data source. Thus, we would choose the representation in which the information loss is minimised.

3.2. Maximum Information Perservation Flag Aggregation

After measuring the information, we can apply to the process of aggregation when the maximum amount of information is preserved under aggreagtion.

In a typical aggregation scenario or computation of derived statistics, we have:

$$S_{agg} = f(S_1, S_2, \cdots, S_n) \tag{1}$$

Since we are restricted to a finite list of available symbol representing different information sources and only one symbol can be used; the loss of information is unavoidable under aggregation. The goal here is to choose a function f which conveys maximum information.

Using entropy, we can quantify the information associated with each data source. The equation then becomes:

$$S_{agg} = F(\max(H(S_1), H(S_2), \cdots, H(S_n))) \tag{2}$$

Where F represents the conversion of the entropy to the character symbol representing the data source. This allows us to preserve as much information as possible under aggregation.

4. Application

The value of cross-entropy itself is not very useful for users, thus we have created what we call the information weights for users to utilize for modelling and subsequent analysis.

4.1. Information Weight

The weights of different information source can be calculated based on the Kullback-Leibner as follow:

$$\omega_i = \begin{cases} 1/(1 + D_{\mathrm{KL}}(P \| Q_i)) & \text{if } D_{\mathrm{KL}}(P \| Q_i) \neq 0 \\ 1 - 1e^{-5} & \text{if } D_{\mathrm{KL}}(P \| Q_i) = 0 \end{cases}$$

Essentially, the greater the loss of information the less the weight. This is in corcordance with our expectation that the lower the information quality, the evidence it provide is less and thus taking less weight in the model or analysis.

5. Conclusion and Further Work

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_flag