American Journal of Agricultural and Biological Sciences 2 (1): 23-30, 2007
ISSN 1557-4989

Data Quality and Indicators

[1]Jorge Matute and [2]A. P. Gupta
[1]Private Consultant, Guatemala, c/o Section 411, PO Box 2-5289 Miami, Florida 33102-5289, U.S.A.
[2]Department of Soil Science, CCS Haryana Agricultural University, Hisar, India, 125004

Abstract: This study highlights the importance of collecting good quality data from multidisciplinary studies. Bias in data may be the result of instrument inaccuracies, imprecise data recording techniques, inaccurate data entry to computers or inappropriate statistical analysis and presentation. Recommendations for good data quality control are given. Different types of data are discussed: raw data, simple indicators and complex indicators. It is shown how measurements from the components of multidisciplinary systems can be combined to form complex indicators and a specific example is given using Z-scores and dot charts. Finally the accumulated effect of bias in the individual component measurements upon the combined indicator is shown.

Key words: Data quality, indicators, bias

INTRODUCTION

The quality of experimental or survey results depends upon the quality of their associated data. In research programs, whatever their size, the maintenance of data quality is a continuous problem. Staff must be trained both to collect accurate data and to enter them into computers in a precise manner. If this is not done then data with errors will lead to faulty analyses and wrong conclusions and decisions. In many field studies, program managers and researchers are very confident about their data collection staff, trusting their fieldwork and data-entry processing without question. But everybody is susceptible to making mistakes, so data quality should not be taken for granted. Many publications have stressed the need for good quality data[1,2] and the implications of poor data recording[3,4]. These stress the large effects in even quite small univariate studies. But in large multidisciplinary studies where raw data, simple and complex indicators are involved, the accumulated effect of initial collection of poor data will have even greater implications on the results and conclusions. Here we show how biased data are generated and can have serious implications on complex indicators.

Data quality: Researchers and program managers need to obtain precise estimates of the effect of some intervention in order to make a decision or a recommendation. The estimate of the intervention effect should reflect as closely as possible what would be determined if the whole population could be measured. When the collected data on which the estimates are to be based do not accurately reflect reality, they are said to be *biased.* Biased data will provide biased estimates and may lead to wrong decisions. For example, if a dataset implies that a community under a health improvement program no longer has health problems (or that there are minimum problems), then the decision may be taken to start a similar health improvement program in other communities or to focus on other domains. If the first community health status has not improved to such a positive extent then the new health programs will be waste of resources. Another example of biased data is when someone tries to evaluate the effect of a program to determine to what extent children are being helped and motivated to attend school. From collected data one can decide whether the program has been successful. If the sample of children that was taken was not random, if the sample size was not large enough to include children from all grades or if there was a systematic error in data collection the data are likely to be biased and may distort the real effect.

Data can become biased from several different sources and processes: These are:

* The instruments used for measuring data
* The field personnel who gather the information

Corresponding Author: J. Matute, Private Consultant, Guatemala, c/o Section 411, PO Box 2-5289 Miami, Florida 33102-5289, U.S.A., Tel: +502 471 5826, Fax: +502 471 5826

* The entry of the data to the computer

The summaries of the data, whether they are biased or not, can themselves introduce a degree of misrepresentation to the results. This will come from:

* The statistical analysis
* The presentation of estimated values

Each of these is dealt with here.

Instruments for measuring data: Instruments may include simple tools like questionnaires or more sophisticated equipment such as automatic data loggers. Whatever their degree of complexity they must all be used with the same consideration for precision. Sometimes an inexperienced researcher may think that just to use a questionnaire with simple questions is enough to obtain the desired data. If the questions in the questionnaire were not validated or did not include the vital variables to answer the questions or objectives of the study then the resultant data will at best be incomplete and at worst, biased. Before going to the field to collect the data, it is important to test if the questionnaire measures what is required. A simple example of this is the language to use. For example, in some rural areas in Guatemala, the word diarrhea does not exist, instead there are many different local names. An inexperienced researcher may use the word diarrhea and get answers that will not be measuring the reality of this indicator.

Electronic instruments must be set accurately. Simple weighing scales and automatic data loggers, for example, must be calibrated at the start of a study and must be checked regularly for their reliability of measurement. If the instruments are not set correctly initially, then they could be including some extraneous measurement together with the effect that is required. This is known as instrument bias.

The field personnel who gather the information: The process of training field personnel is important, because this is the first major influence on data quality. It is essential to:

* Establish the quality maintenance procedures during selection of sampling units,
* Train the field personnel to handle the instruments in the same way. If this standardization is not achieved, they may interview in different ways,

During the field collection of data, it is important to monitor the quality of:

* The field personnel making the interviews and measurements,
* The selection of sampling units,
* The handling of instruments

All of the above are potential sources of bias, that are important to remember and to control in order to minimize the bias as much as possible.

Specific types of bias are generated from biased samples, which have no randomness in the selection of the sampling units and from systematic errors which are usually made without the knowledge of the data collector. There are different possibilities for making systematic errors:

During data capture in the field an interviewer may interpret a response in a different way from the rest of the interviewers, or may code the information differently from the rest of the field workers.

Machinery may not be set accurately and may add (or subtract) a constant value from the true measurement

The entry of the data to the computer: Once the fieldwork is done, or even at the same time, the process of entering the data to the computer begins. This should aim to enter the data exactly as it was recorded in the field. In other words, it should be like xeroxing the field work into the computer. But, unless the field instruments are read directly by a computer, there is always human error involved in the process. During the data entry process, the data-entry staff does their best to maintain accurate data entry. They enter data very fast, depending on the clarity of the data being entered. However, data entry errors may still be made. This is understandable**,** but to ignore them because there is no quality control is not acceptable. If the conclusions from an analysis of the data clearly point to a nonsense result, then the data entry errors may be spotted, otherwise the errors will be hidden and the information misleading. The result of this is that the researcher will have to pay for his mistake by spending a lot of time in finding the errors and doing the analysis again.

Different procedures to control the problems of data entry errors have been developed. The most common are:

Data entry → manual check→ errors fixed: Here there is a likelihood that errors will remain. This is a common method among people with little or no experience in how to deal with this problem. The data are entered once into the computer and an output produced. The output is checked manually with the original documents. This is time consuming and does not guarantee correction of all the errors (even if the checking is made by more than one person).

Data entry → sample data→ Fix errors in sample: This is not a recommended procedure. Its main purpose is to understand how good the data are; in no case is it to correct the errors. The data are entered once and a random sample of data (cases, individuals) is chosen to check for errors among them. An estimate of the errors due to data entry is obtained with this sample. However, only the errors in the sample are corrected, but not those in the rest of the data. A study may have 10,000 children and only one data entry has been paid for. If a random sample of 100 children is selected from the typed data to check the data entry quality shows that 10% of the data entered are wrong, these 10 errors may be corrected but the remaining estimate of 990 errors will not be. This second option for checking data is not viable.

Double data entry → compare the two entries → fix errors in one of the files: This is one used by most professional organizations with experience in dealing with data entry. It consists of entering the data twice; and then the two files are compared by means of ***validation*** software. In contrast to the first option, this one has the disadvantage of having to enter all the data twice which implies twice the time and possibly twice the costs for data entry. However, by using a computer the comparison between the files is done in minutes, instead of having one or more persons checking manually for errors. The time cost consumed in manually checking is invested in a second entry and which usually is faster than the process of manual checking. This third option will be more cost effective than the first one, with the benefit of zero errors in the data or a very small amount of error.

Double data entry → compare two files → fix errors in both files → compare again → Fix errors again, etc. until no errors are found

This is an extension of the third method. Experience has shown that after comparing the files and correcting the errors - usually they are in only one of the files - this is not sufficient to eliminate all the errors. For example, if a file is very large it may be necessary to take a break in the checking process and this may result in loss of the last variable or case that was worked on. So, some of the original errors remain. So one way to deal with this is to correct the files repeatedly until no further error is found. So far, this option has proved to be the most efficient of the four, with the guarantee of 0% errors in data entry. The procedure requires no more appreciable amount of data personnel time or cost than the third procedure.

The implications of organizing good data collection and handling procedures are that there will be more work to do, professionals to contract and thus a direct impact on the budget. This may be unwelcome to the administrators. But the greater investment will be cost-effective. There is nothing more inefficient than running a large study and not being able to use the data because of inherent biases.

The statistical analysis: The development of software and the introduction of statistical software tools for common use have resulted in more and more people using statistical methods. However, bias may arise when these statistical tools are used in the wrong way. For example when parametric methods are used without checking their assumptions or when a hypothesis test is inappropriately used. Problems may arise where the distributional properties of the data demand that a transformation of the values is made before analysis. If this is not done then the results of the analysis can easily lead to invalid conclusions.

But besides inferential analysis, there are descriptive analyses used by almost all agencies to evaluate their programs. In any project in any country, there may be bias when an analysis of a survey is based on a simple random sampling design rather than the complex sampling designs that may have been used, such as stratified-cluster sampling. If the data analyst has not been informed about the correct design structure, the results could range from a little bias which may not be very influential to a large one which may lead to misleading results depending on the variable or indicator being analyzed.

Misleading results are likely to have poor external validity so professional statistical advice is recommended to all researchers at all stages from design of their study through to statistical analysis.

The presentation of estimated values: The sampling process provides an estimate of a population value. However the estimate is not a single value, but is a value within a set of bounds, limits or thresholds dictated by the variability in the data and rules according to probability levels. This range is the Confidence Interval. It is important for any technical report to give confidence intervals together with the estimated parameter (which may be a mean value) and an estimate of its variability, say its standard deviation. This is very important when dealing with indicators which may be functions of variables, functions of some complexity. It is not sufficiently simply to report the results of significance tests.

A wide confidence interval shows that a lot of variability is attached to the estimate implying that there is a great deal of variability in the data or that the estimation process is not a good one. A narrow interval provides greater confidence in the estimate of the true population value. The presentation of a single estimated value without considering the possibility of a range of values and limits or thresholds to these values is likely to lead to biased or inaccurate recommendations.

Indicators

Indicator definitions: Variables and indicators need to be defined. A **variable is a characteristic to be measured on a subject or unit**, for example weight of a crop, type of soil, gender of children and age of farmers**. A parameter is the unknown population value. Variables are used to estimate parameters.** Using this definition, an indicator can be defined as a parameter with a threshold value(s), that can be used to describe or evaluate a particular system and according to which a decision may be taken.

It is estimated with a variable or a set of variables, which is related to a decision process based on threshold values. For example, malnutrition is defined as the percentage of children with a nutritional status given by a "Z" score (See a more detailed definition of the Z score ahead on the section "Construction of complex indicators") on weight for age under 2 standard deviations; measure the malnutrition in a given region and then place a threshold on it, like decreasing it by 10%.

There are indicators that are defined by a simple variable, for example age of a farmer; these indicators are referred as simple indicators. A variable may be a part or a whole component of the system. Thus, yield is an indicator and it can be defined by several variables such as plant population, total number of tillers, total number of productive tillers, number of grains per ear head and test weight (1000-grain weight). In this case, yield may be defined as a complex indicator because in order to get it, it is necessary to estimate different variables and to put them together in a particular way. In other words these complex indicators are functions of simple indicators. Other examples of complex indicators are increase in yield, soil health and malnutrition.

There can be several ways to classify indicators. Based on their measurement scales, indicators can be classified as qualitative for example socio-economic status, gender, soil health or quantitative for example age of a person, yield increase, benefit to cost ratio, literacy level (for example if the person reads or not).

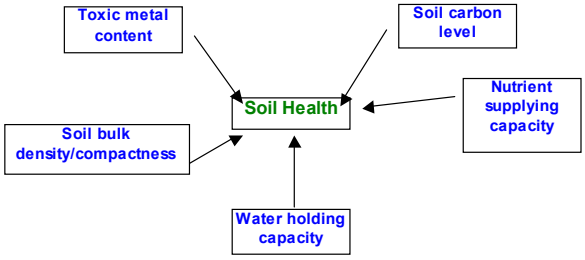

Fig. 1: Main indicators contributing to soil health

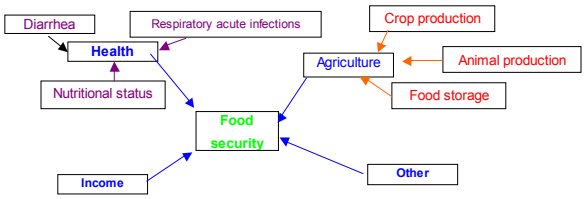

Fig. 2: Main indicators contributing to food security

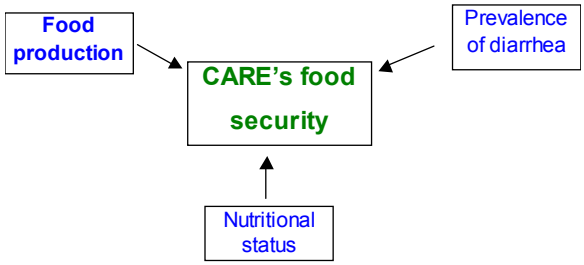

Fig. 3: Some indicators contributing to CARE's Food Security

Construction of complex indicators: Soil health as an indicator is dependent on the combined effect of several other indicators. Figure 1 shows the five main indicator groups which contribute to soil health. It may be necessary to weigh each indicator group differently. The weights can be assigned based on the extent to which they affect the complex indicator. For example toxic metal content will influence soil health to the extent it will kill any crop thus this indicator will get more weight than the others.

Another important indicator is food security. This is the combination of several other indicators on health such as nutritional status, presence of water supply systems, diarrhea, disposal of fecal excrements, agricultural indicators such as crop production, food storage, animal production, as well as income and other variables. This is shown diagrammatically in Fig. 2.

A real example using data from the Food Security Program from CARE International in Guatemala is demonstrated here. CARE had not worked on all the areas related to food security, but wished to measure the

effect of the different interventions on a single food security indicator. Therefore simple indicators related to the interventions were chosen to define food security. These indicators and their relationship with the main and complex indicators - food production, prevalence of diarrhea and nutritional status are shown in Fig. 3.

Nutritional status can be measured by different indicators related to anthropometry, the one used in this example is in itself a complex indicator called a "Z-score" on weight for age (WAZ). It is a standardized value of children's weights, compared with the mean and standard deviation from a "standard population". Z-scores have a normal distribution with a mean of zero and standard deviation of 1. Malnutrition, in this case underweight, is defined as being lower than 2 standard deviations below the mean value. Prevalence of diarrhea is a simple indicator, created by a single qualitative variable - yes, no - that is yes if a child had diarrhea during the last 15 days.

Food production is a complex indicator formed from six other indicators:

* Weight of maize production
* Weight of beans production
* Percentage not lost in maize storage
* Percentage not lost in beans storage
* Animal production - Existence of small animals at the household
* Existence of a garden at home (with plants and vegetables rich on Vitamin A and iron).

So, food production is a complex indicator formed with three different types of indicators. It was decided that each of the six indicators that leads to food production should have the same weight, so there was no one indicator more important than the other. This decision made it easier to find the solution, which was to transform all the indicators to the same scale, here to standardize them to Z-scores:

$$Z = \frac{\text{observation} - \text{mean}}{\text{SD}}$$

where "mean" is the arithmetic mean of the distribution and *SD* is its standard deviation.

However, by using the mean from their own distributions, all transformed indicators (Z-scores) will have zero as a mean (by definition). When these transformed indicators are ultimately combined or averaged the result may be zero; certainly a loss of information is likely to occur. Therefore, instead of calculating the Z-scores with the means from their own distribution, they were calculated using the proposed goals as means. Thus the indicators have non-zero means and their values reflect how far from the goals the indicators are (the goals being the Z-scores equal to zero). Thus,

$$Z = \frac{\text{observation} - \text{goal}}{\text{SD}}$$

Once all Z-scores for each indicator were obtained, then a mean from all of them was calculated. So to obtain the food production indicator, it was calculated as the mean of all the Z-scores from the six indicators, using different goals appropriate for each indicator,

$Z_{\text{Food Production}} = (Z_1 + Z_2 + Z_3 + Z_4 + Z_5 + Z_6) / 6$

The method of getting the food security indicator was the same as that for food production. But to have all the indicators pointing to the same direction, instead of calculating Z-scores for diarrhea, the Z-scores for NO diarrhea were calculated. So, food security was defined as:

$\text{Food security} = (Z_{\text{food production}} + Z_{\text{WAZ}} + Z_{\text{No diarrhea}}) / 3$

Then the mean of Food Security was estimated as follows:

$$\text{FoodSecurity} = \frac{\sum_{i=1}^{n}\left(\frac{Z_{\text{food production}} + Z_{\text{waz}} + Z_{\text{NO diarrhea}}}{3}\right)}{n}$$

These three indicators are calculated for each of *n* individuals and then added and then divided by 3. All n values are then added and averaged over the n individuals. The same applies for all the other indicators used to estimate food security.

Graphical representation of a complex indicator: Graphing a complex indicator may be considered difficult to do, if not impossible, but it can be done in a simple way, similar to its estimation. Dot charts[5] can be used and because the food security program from CARE had a baseline and the design was a follow up, this format is ideal for demonstrating the effects of the indicators.

In Fig. 4, it is easy to see that by using Z-scores the idea of being well or better is attained by having larger positive values. Remember that diarrhea prevalence and losses in storage were reversed to ensure that improved conditions related to increased values.

There are four panels in the graph. Panel "1" shows how the six indicators that form food production change from the baseline to the mid term appraisals. Indicators on gardening and animal production did not change. For crop production, only maize production had improved but the bean production is almost constant. Finally the indicators on NO loss at storage improved, meaning that the losses were less at midterm than at the baseline. The food production indicator is shown in

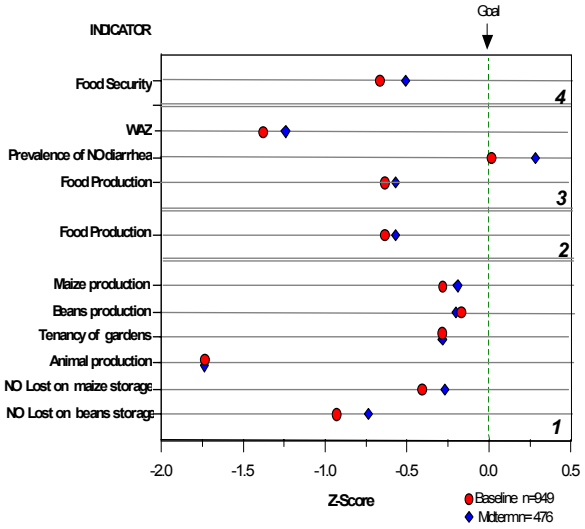

Fig. 4: Food security indicators

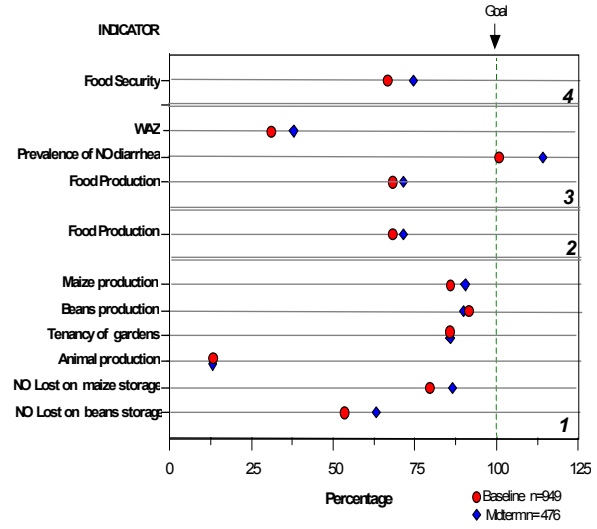

Fig. 5: Food security, scale in percentages

panel "2", which has a little improvement since baseline (because of the positive change in maize production). In panel "3" are the three indicators that form the Food Security indicator. Prevalence of NO diarrhea was the one that improved more than the other two. WAZ also moved in a positive direction (meaning that children improved their nutritional status). Together, the movement from these three indicators leads to the overall effect on food security that is shown on panel "4" (a change of 0.16 SD, meaning an improvement of quality of life for these people).

Presentation of this information to policy makers: Although the above graph is very useful and easy to interpret for someone with a little knowledge in reading graphs and understanding of Z-scores, it may be difficult to decode for others.

The first step to make the graph intelligible is to change the scale on the "X" axis. Percentages can be presented instead of Z-scores. To do this, the Z-score of zero is defined as 100% (achievement of the goal). Under the normal curve, within 1.96 standard deviations of the mean value, there is an area of 95%. Taking the value of –2 standard deviations as the zero value for percentage achievement of the goal, then a simple transformation gives:

Z –score	-2.0	-1.5	-1.0	-0.5	0.0	0.5
Percentage (of achievement of the goal)	0	25	50	75	100	125

So, the transformed Fig. 4 will appear as Fig. 5. This may still appear somewhat complex but it is the clearest representation of a multidimensional indicator that can be achieved. Other, simpler, presentations may be attempted such as bar charts. These however are not recommended because the differences will be less apparent.

Weighting to construct a complex indicator: For the above example, the complex indicator related to food security was constructed assuming that the three component indicators had the same influence. Thus,

food security = $(Z_{\text{food production}} + Z_{\text{WAZ}} + Z_{\text{No diarrhea}}) / 3$

Some indicators may be considered more important or relevant than others (may be due to the way the intervention was planned). Therefore, they should be weighted differently. For example:

IND 1 = $(Z_{\text{food production}} + Z_{\text{WAZ}} + Z_{\text{No diarrhea}}) / 3$ (*This is the original)*

IND 2 = $(2Z_{\text{food production}} + Z_{\text{WAZ}} + 3Z_{\text{No diarrhea}}) / 6$

IND 3 = $(Z_{\text{food production}} + 4Z_{\text{WAZ}} + Z_{\text{No diarrhea}}) / 6$

IND 4 = $(3Z_{\text{food production}} + Z_{\text{WAZ}} + 9Z_{\text{No diarrhea}}) / 13$

IND 5 = $(Z_{\text{food production}} + 3Z_{\text{WAZ}} + 3Z_{\text{No diarrhea}}) / 13$

Figure 6 shows how the different weightings affect the final value. The example shows data from the midterm evaluation only. So, in constructing a complex indicator the way that its indicators or variables are weighted are critical. But the choice of weights themselves is difficult. This is an issue that should be addressed preferable together with a statistician while designing of the study or at least before the statistical analysis is started.

Indicators and bias: Indicators and bias have been described in sections 2 and 3, but how does bias affect

Table 1: Results of simulated bias process

Indicator	Unbiased	Biased	Bias (%)
Food production	-0.571	-0.580	1.6
Food security	-0.509	-0.630	19.2

Table 2: Bias and design for indicators analyzed according to an incorrect design

Indicator	Bias (%)	Design effect
Baseline evaluation		
WAZ	3.8	5.7
Prevalence of NO diarrhea	-64.2	1.3
Food production	6.6	8.5
Food security	5.1	6.4
Midterm evaluation		
WAZ	-4.0	1.3
Prevalence of NO diarrhea	1.0	2.0
Food production	2.8	7.1
Food security	2.2	1.7

an indicator? To show how bias could change an indicator, two examples are presented here.

The first example shows how a systematic error, mainly coming from the field, could change the results. Assume that the fieldwork was done by ten people and that one of them was not standardized and made a systematic error. To do this we randomly biased 10% of the results from the midterm evaluation, on the "food production" indicator, by subtracting 0.5 standard deviations from the standardized indicator. This food production indicator is part of the "food security" indicator so the bias is reflected in both indicators as can be seen in Table 1. In this case bias did have a small effect on food production, but it had lot of influence (19%) on the final and most important indicator - food security, which would lead to erroneous conclusions and wrong decisions regarding the intervention.

The other example reflects one of the most common mistakes while analyzing data and which is to forget that a complex sampling design was used and to analyze the data according to a simple random sample format. The sampling design used to evaluate CARE's food security program was a stratified – cluster sampling (two stages with probability of selection equal to size). So, it is not a simple random sample. To show how the indicators may be biased by not analyzing the data correctly, the indicators "food production", "prevalence of no diarrhea", "WAZ" and "food security", were analyzed assuming the incorrect simple random sample design. The bias and design effect for the indicators at the baseline and midterm evaluations are presented in Table 2. A design effect is the influence given by a sampling design to the variance. A simple random sample is defined as not having an

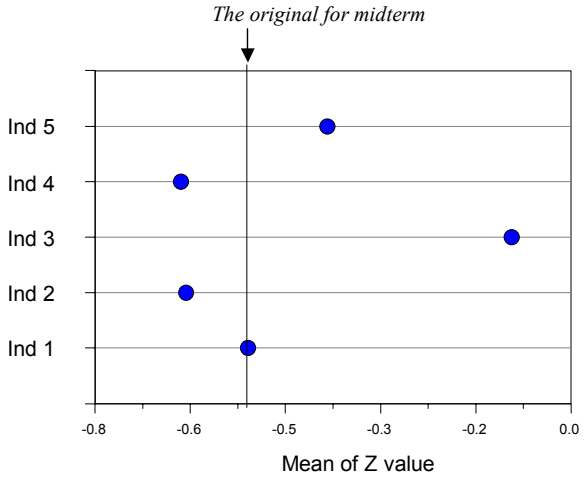

Fig. 6: Effect of different weights on simple indicators to estimate a complex one

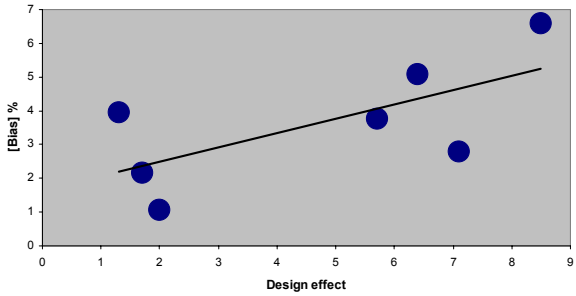

Fig. 7: Bias and design effect

effect on the variance, any other design will have an effect on the variance (which could be positive or negative, positive meaning to reduce the variance and negative to increase the variance). A way to calculate the design effect and at the same time to show the meaning of it, is given by the following equation:

$$\text{Design Effect} = \frac{\text{Variance from a complex design}}{\text{Variance from a simple random sample design}}$$

From the above example, values that have a design effect less than 2 are related to bias in a non predictable way, while values above 2 demonstrate a direct relationship between bias and design effect (Table 2).

Bias may be small or large. If it is large, then certainly wrong conclusions would be stated. If it is small, it may not affect the results and will remain close to a threshold. It is never known for certain if the data are biased, but likely causes of bias should be avoided as much as possible in order to support good data analysis leading to reliable recommendations and decisions.

CONCLUSION

Data quality is essential to support good analysis and estimation procedures both for simple and complex multidisciplinary surveys and experiments. Bias in data comes from several different sources and control of these can be maintained with a rigorous checking procedure. Equipment and personnel both are sources of error and thus bias in data.

It is essential to decide about the survey proforma in consultation with a statistician so that all the possible facets are taken care off. The design to be followed should also be clear. The field personnel should be well aquatinted with the filling of the proforma. There must be some cross checks on the field information to know the bias in the data collected. After taking a decision from the collected data, it is also essential to revalidate the results in a small area to confirm the findings.

Bias in data may be small or large. If it is small then resultant bias in indicators and their analysis should be small. Indicators formed from collected data may be simple or complex, but provided data quality is good then the results of analyses of complex indicators should be unbiased. It is important to stress that decisions are made from measurements on indicators and if these indicators are biased the decisions may be wrong and even worse, they may lead to non-results or unexpected and misleading effects.

ACKNOWLEDGMENTS

This article was made as an aside product during the UNIQUAIMS project, which was a concerted action between many European and developing countries, working on renewable natural resources programmes; lead by The Statistics Department at Rothamsted Research, Harpenden (AL5 2JQ, UK), under the direction of Janet Riley. We thank Dr. Riley for her support and guidance during that time.

REFERENCES

1. Panse, V.G., 1954. Estimation of Crop Yields. Rome, Food and Agriculture Organization.
2. Dyke, G.V.D., 1988. Comparative Experiments with Field Crops. 2nd Edn. London, Griffin.
3. Preece, D.A., 1984. Biometry in the Third World: science not ritual. Biometrics, 40: 519-523.
4. Fielding, W.J., 1992. Damage assessment by eye: Some Caribbean observations. Field Crops Res., 30: 183-186.
5. Cleveland, W., 1985. The Elements of Graphing Data. California, USA, Wadsworth.