Large, secondarily collected data in biological and environmental sciences

T. N. C. Vidya

Large, secondarily collected data are often used in the biological and environmental sciences in order to gain broad insights. However, if the data are collected by untrained or unskilled people who do not appreciate the significance of the final dataset, the dataset may become worthless. We need to examine the data collection process in order to use only data of high quality.

The newspapers inform us that the integrated all-India tiger census has just begun[1,2]. This census will enable data collection on tigers and their prey-base on an unprecedented scale. Thousands of volunteers and forest staff across the country are expected to record everything from tiger sightings to herbivore pellet counts to the percentage of different tree, shrub and grass species in vegetation plots. But how reliable are such large-scale datasets collected by volunteers or relatively untrained staff? The nature of data collected as well as the ability and honesty of the data collectors come to mind as possible determinants of the reliability of such datasets. Data should be simple and non-subjective for reliable collection by motivated volunteers or staff. The Centre for Tropical Forest Science (CTFS) is a global network that monitors forest plots, some of which are entirely censused by volunteers. Measuring the girth at breast height (GBH) of tagged and numbered trees is a simple and repetitive task. Even so, such measurements taken by a large number of people are useful only if the exact heights at which trees are to be measured are marked on individual trees (they are in many plots). On the other hand, assessing the percentage dominance of a specific grass species is both a subjective and an inordinately skilled job. There are probably a handful of experts in the entire country who can identify grass species (other than the most common ones). Therefore, despite good intentions, such a survey will lead to a lot of meaningless data, although there might be some subsets of data that are more reliable.

This census is only a recent example, but many such large datasets come to mind that could be fraught with problems of quality. There are numerous weather stations situated in remote areas of the country, where temperature and rainfall have to be measured and entered manually every day. These are despatched to offices up the hierarchy until figures for districts and states are obtained. But do we know in how many of these remote areas weather parameters are recorded with some degree of accuracy every day? (This may be more relevant to past data rather than present-day data, given the availability of advanced recording systems nowadays.) Although the nature of the data collected in this case appears simple, local staff may not be literate enough to read Arabic numerals, deal with decimals, or realize that a piece of equipment is malfunctioning. Second, staff may not lack the ability to perform the task, but may feel that the task is not important enough to warrant complete honesty. For instance, how often would they hesitate to write down, say, the previous day's values when a recording was missed? Similarly, various data related to crop productivity are collected across the country, but how rigorously is this done? Are those who actually collect these data from crop fields trained in assessing the error in their measurements? Do they even care about errors? (This applies to researchers too, collecting any kind of data.) Lack of such rigour in reporting, I suspect, is possibly even more marked in the health sector. Statistics about the incidences of various diseases across the country are often displayed, but how exactly are these data collected and who actually records them? Are at least a majority of the incidences recorded?

Should we be using these kinds of large datasets collected by insufficiently trained or motivated staff or volunteers in scientific analyses? Is it, for instance, meaningful to model the spread of a virus, using prevalence data available from a state? The problem with big dataset science is that it is only as good as the weakest link, and, in a country with huge numbers of inadequately trained people, there are many weak links. A wrong GPS reading (even those with a little training have been observed to wrongly note down a degrees–minutes–seconds notation as a degree–decimal–degree GPS reading or vice versa) can extend the range of a species. Mistakes in the decimalization or entry of a few numbers can change the average rainfall of, or the average incidence of malaria in a taluk. Such errors abound in the absence of rigorous checking of data at every level and, once the data are compiled at larger scales, such errors will become impossible to pinpoint. One might be tempted to think that different kinds of errors in large datasets would get averaged out, but I do not know of any evidence that supports this. Consistent biases can arise from various sensory and mental perceptions, ranging from biases in transcriptional errors (due to the similarity of certain letters or numerals) to those in expectations (which can strongly shape observations)[3].

There are many studies based on personal interviews of stakeholders in rural areas that are carried out by volunteers with the help of local translators and the way questions are posed can lead to the desired answer[4], leading to biased outcomes depending on the interviewers' or translators' perceptions. Moreover, if the data are not collected primarily for a scientific study, there could also be vested interests in biasing data (for example, forest cover, disease incidence or human mortality).

India is by no means unique in facing this problem of low quality, large datasets. In a scientific world dominated by the pressure to publish, there is, unfortunately, a temptation to use large secondarily collected datasets, as they yield high-impact papers in relatively short periods of time. Even graduate students across the world in some fields seem to be increasingly delegating the job of

primary data collection to technicians and assistants. However, the farther we move away from primary data collection, the more difficult it is to be rigorous about it. Moreover, complex statistics and models may dazzle some into not seeing the poor-quality data beneath. While we may not be unique in this, one important reason why I think we should be wary of large-scale secondarily collected data is that we as a culture do not like to admit that we do not know something, however trivial. We are also not a people who can easily refuse something that is asked of us. Therefore, asking someone to fill out a datasheet without strict supervision is likely to lead to all columns being filled out, even if with gibberish, rather than an honest account of missing or questionable data. The big picture is rarely explained to those actually collecting data, and the end-users of the data often have little connect with field conditions.

Sting operations such as that conducted recently by *Science*[5] may reveal outright fraud, but low data quality is more insidious and difficult to assess, and depends on the lack of ability or basic honesty of people collecting the data. If overall honesty is correlated with the levels of corruption, we as a country have much to be worried about (see ref. 6, although they are also large secondary datasets). I strongly suggest that we be vigilant about data quality, use only secondary data that are simple and objective, take the trouble to examine the entire data-collection process first-hand to understand what the shortcomings of the specific dataset might be, and try to explain the larger significance of the dataset to those involved in data collection.

9. Bennur, S., *The Hindu*, 16 December 2013.
10. Maramkal, M. B., *The Times of India*, 15 December 2013.
11. Wiseman, R., *Quirkology: How We Discover the Big Truths in Small Things*, Basic Books, 2008, p. 336.
12. Lynn, J. and Jay, A., *The Complete Yes Minister*, BBC Books, 1989, p. 514.
13. Hvistendahl, M., *Science*, 2013, **342**, 1035–1039.
14. Transparency International, Corruption Perceptions Index 2013; http://cpi.transparency.org/cpi2013/results/ (accessed on 22 December 2013).

T. N. C. Vidya is in the Evolutionary and Organismal Biology Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560 064, India. e-mail: tncvidya@jncasr.ac.in

Smile with Science

By – Santosh Kumar Sharma
e-mail: santosh_ujj@yahoo.com