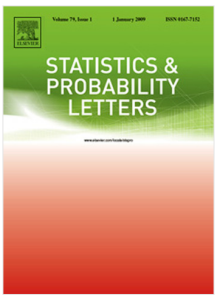

When small data beats big data

Nicole Augustin, Julian Faraway

PII: S0167-7152(18)30076-2
DOI: https://doi.org/10.1016/j.spl.2018.02.031
Reference: STAPRO 8153

To appear in: *Statistics and Probability Letters*

Please cite this article as: Augustin N., Faraway J., When small data beats big data. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.031

When small data beats big data

Nicole Augustin and Julian Faraway

Department of Mathematical Sciences, University of Bath

Abstract

Small data is sometimes preferable to big data. A high quality small sample can produce superior inferences to a low quality large sample. Data has acquisition, computation and privacy costs which require costs to be balanced against benefits. Statistical inference works well on small data but not so well on large data. Sometimes aggregation into small datasets is better than large individual-level data. Small data is a better starting point for teaching of Statistics.

Keywords: Big data, small data.

1. Introduction

Big data is justifiably a major focus of research and public interest. Even so, small data is still with us. The same technological and societal forces which have generated big data have also generated a much larger number of small datasets. At first glance, more data would seem to be clearly better than less data. All things being equal, this is true. In practice, obtaining more data will involve additional costs of various kinds and will complicate the analysis. In the real world of fixed budgets, there are trade offs between quality and quantity. Sometimes small data will beat big data and reach the right conclusions faster, more reliably and at lower cost. In this article, we present a variety of situations where small data will be preferable. For related discussion in this same special issue, see Secchi (2018).

2. Wider Meaning of Big and Small data

The term "big data" means different things to different people. Statisticians tend to think of "big" in terms of size, either many cases or many variables or both. Yet the term has taken on a wider meaning to the public with "big" also refering to the extent, impact and mindshare of the phenomenom. Statisticians have had to adapt their communication to this wider definition. This is now accepted and understood. It is perhaps less well-known among statisticians that "small data" also has a wider meaning in the business community as a reaction to big data. As with big data, the definition proves elusive but we attempt a contrast. Big data deals with the large, observational and machine analysed. Small data results from the experimental or intentionally collected data of a human scale where the focus is on causation and understanding rather than prediction. See the book, "Small Data", by Lindstrom (2016). Given the hype surrounding big data in the business world, it is refreshing to see some recognition for the virtues of small data.

3. Quality beats quantity

In 1936, the popular *Literary Digest* magazine ran a poll of its readers to predict the result of the US presidential election. 2.4 million people responded to the poll with 57% favouring Alfred Landon and 43% for Franklin Roosevelt. In the election, Roosevelt beat Landon in a landslide victory, 62% over 38%. That same year, George Gallup was getting started with his polling organization. Using a sample of just thousands, Gallup predicted a Roosevelt victory with 56%. How could the small data of just thousands beat the big data of millions? Any estimator is vulnerable to bias and variance. Readers of the *Literary Digest* had the discretionary income to spend on a magazine and were typically wealthier than the general population in a time of severe economic depression. The bias was not mitigated by the large sample size. Gallup's small sample would have been subject to greater variance, but this was a far less serious problem than the bias. See Freedman et al. (1998) for details.

In the previous example, we can clearly see how the bias arose and might see ways in which this could be avoided or mitigated. However, bias in big data is sometimes more subtle and less obvious although the consequences can still be severe. Consider the following example from Meng (2014). Suppose we have the choice of taking a small probabilistic random sample from the population or a large adminstrative sample with fraction f_a of the population. We are interested in estimating a quantitative value such as average income. The mean squared error(MSE) is the variance plus the bias squared. For the probabilistic sample, the variance is s^2/n where n is the sample size and s^2 is the population variance. If we do the probability sampling correctly, there will be no bias.

The administrative sample will be large but bias may be a problem. Let r be the correlation between the true response and the probability of a response being available or being observed. The squared bias is given by:

$$r^2((1-f_a)/f_a)s^2$$

The MSE for the administrative sample will be almost equal to this since the variance will be negligible due to the large sample size. Lets say we have only a small random sample of 100 and suppose that the correlation r is a rather small 0.1. Under these circumstances, by comparing the two MSEs, we see that we need $f_a > 0.5$ for the administrative sample to have a lower MSE. This is quite surprising since the random sample is so small and the correlation weak. This illustrates the surprising power of a quality small dataset.

More widely, researchers often prefer a small sample collected in controlled experimental conditions to a large observational sample of unknown provenance. Where an inference of causation is desired, quality beats quantity in data.

4. Cost

The real world is constrained by resources and data has a cost. In parametric inference, accuracy in estimation increases at a $\sqrt{n}$ rate. Although some economy in scale is sometimes possible, the costs of data collection usually increase

at rate n. We will want to minimize an expression of the form $an+b/\sqrt{n}$. Some utilitarian calculus is required to choose appropriate values of a and b but there will be an optimum sample size beyond which more data collection cannot be justified. In practice, people find it difficult to quantify a and b, so the choice of n is not made exactly. Nevertheless, people are aware of this tradeoff. Power calculations are another example of such economies in action. Researchers understand the need to keep the sample size no larger than necessary.

Acquisition costs for data are familiar but there are other types of costs that need to be considered. Computation and privacy costs can matter. As before, we might suppose that the accuracy of the inference improves at best with rate $\sqrt{n}$. For nonparametric procedures, the improvement is at a lower rate. For large datasets, we are more likely to resort to nonparametric procedures due to the difficulty in scaling up parametric methods. Let us consider the computational costs. At best, these will increase at rate n but sometimes it will be much worse than this. A statistical procedure that involves a matrix inversion will increase at rate n^3. A problem that is NP-hard will be even worse than this. Some applications, on the internet or in online control, require a fast answer - perhaps almost in real time. In such circumstances, there is a limited computational or time budget. One might have to choose between a simplistic analysis of a large dataset or a sophisticated analysis of a smaller dataset. The latter may be the better choice. See Chandrasekaran and Jordan (2013) for more on this.

A variety of externalities can be associated with data. Privacy is a major concern in some situations. Obtaining informed consent and considering the potential loss of privacy for subjects in a study can be expensive. If questions can be answered with a smaller dataset, we prefer this to threatening the privacy of a much larger number of individuals.

Scientific research, as a search for the truth, is conducted without regard to time or cost. We would be prepared to wait any amount of time or bear any cost in order to obtain the truth. Outside of utopia, we must balance the cost of the data we use against the benefit we hope to obtain. We must make the

best of limited resources.

5. Statistical inference works better on small data

Most textbooks and learning materials in Statistics concentrate on data of a modest size. This is partly from convenience the resulting inferences have at least some uncertainty illustrating the essence of the methodology. Big data is problematic for standard statistical inference.

Although substantial theoretical effort has gone into asymptotic analyses, these are of little practical use when n becomes large. For any finite parametric model, confidence intervals become extremely narrow and p-values become very small indeed (barring the unlikely situation that the null hypothesis is actually true!). Bayesians fare little better as the prior is swamped and the likelihood dominates with similar, all too sharp, inferences. One can use non-parametric inferences to grow the parameter space at a sensible rate to avoid some of these problems. But, even with this, the inference become far more certain than common sense would allow. Machine learners have tended to avoid the problem by not providing estimates of uncertainty.

Uncertainty comes from other sources than unknown parameters. We are not sure what model to use and we are uncertain about the biases and errors in the data. If we were better able to incorporate these in our modelling we would achieve a more realistic result. But this is difficult to achieve.

Small data models are necessarily simple and reflect at least some uncertainty. We know about the dangers of model misspecification. Although the results may not calibrate the uncertainty perfectly, at least the user of the conclusions will understand that they should be cautious and allow for the possiblity that they are wrong.

In contrast, models for big data might be fine for point prediction and classification but we struggle to provide realistic assessments of uncertainty. Also consider the problem that big data suggests a massive number of hypotheses with less protection against the danger of false positive results.

In time, we may learn how to express uncertainty realistically in big data inferences. For now, we might prefer the humility of knowing how we may be wrong to the arrogance of believed certainty.

6. Aggregation

The reduction of large individual data to smaller grouped data may lead to aggregation bias. For example if we are interested in modelling individual relationships in the context of diagnostic models for personalised medicine or if the main interest is in modelling extreme events for example in the context of complex spatio-temporal models for temperature or air pollution. But there are also situations where aggregated small data can be better than individual level big data.

For example in environmental monitoring, estimates of spatio-temporal trends of some environmental indicator, e.g. mean tree health, are of key interest. Often response data are recorded at individual level, but most explanatory variables are available at site level. In tree health monitoring, tree defoliation is recorded at a grid of sites yearly on several individual trees, alongside individual tree age, but explanatory variables such as soil properties are only recorded at site level. The reason is that usually soil properties are homogeneous at site level and they are expensive to measure compared to tree health and age. Trees are all of similar age, because the forest is heavily managed. Aggregation bias is not a problem here, as in this case we are interested in the mean defoliation at a specific location and time. Aggregating the tree level data at site level makes sense here, as it simplifies the model and reduces data (Augustin et al. 2009).

In epidemiological studies physical activity is now often measured by an accelerometer. The newest technology allows to measure acceleration at 50 Hertz or more, for storage the signal is converted into counts and summed over a user specified interval, e.g. 1 minute. At this rate time series of 10080 counts are available per individual if measurements are taken over a week. If the data is used to estimate patterns of energy expenditure in humans, any shorter time

interval for aggregation is unlikely to be useful.

Aggregation has several advantages. It reduces variation and data storage requirements. It is simpler to analyse and often eliminates the privacy concerns associated with individual level data.

7. Teaching

In the past, data was necessarily small and statisticians worked to extract the most value from a little information. Instruction in Statistics was centred around these methods for small data. Sometimes a virtue was made of manual computation on paper or with a pocket calculator to inculcate a deeper understanding of the methodology. Even as computing became cheaper and faster, statistical instructions stuck to small dataset, with a preference for those that could reasonably be printed in a textbook.

As substantially bigger datasets became available with more complex modelling requirements, a new approach was needed. Ideological rigidity in the statistical community left the field open to computer scientists who took the lead in developing methods to deal with big data.

A student who plans a career in analysing data needs to know both the small data world of Statistics and big data world of Computer Science. These two worlds overlap substantially and yet it is often difficult to become skilled in both. There is a rapid increase in master's programs in Data Science which draw in large numbers of students. A large part of the instruction focuses on acquisition, cleaning, manipulation and storage of big datasets. This is valuable knowledge but there is a danger in that small datasets often require only trivial curation. To the data scientist, such small datasets will appear of little interest. Machine learning often performs poorly on small datasets. A student who focuses only on big data skills will have serious weaknesses.

Small data skills are essential to the well-rounded data analyst. This requires an understanding of the principles of Statistics. These principles have not been obsoleted by big data. Many, but not all, of the principles used to analyse small

data apply to big data. Small data is a better starting point for teaching than big data because the skills and ideas can be developed with greater focus and convenience. Starting with big data is a mistake since this can lead to focus on technical skills rather than understanding.

8. Conclusion

Data is not an end in itself but a means to an end. The end is increased understanding, better calibrated prediction etc. More is not always better if this comes with increased costs. Data is sometimes viewed as something fixed that we have to deal with. It might be better to view it as a resource. We do not aim to use as many resources as possible. We try to use as few resources as possible to obtain the information we need. We have seen the benefit of big data but we are now also realizing the extent of the associated damage. The modern environmental movement started in reaction to the excesses of resource extraction. It advocates an approach that minimizes the use of resources and reduces the negative externalities. We believe the same approach should be taken with data: Small is beautiful.

References

Augustin, N.H., M. Musio, K. von Wilpert, E. Kublin, S.N. Wood and M. Schumacher, (2009) Modelling spatio-temporal trends of forest health monitoring data. The Journal of the American Statistical Association. 104(487): 899-911.

Chandrasekaran, V. and Jordan, M.I., (2013) Computational and statistical tradeoffs via convex relaxation. Proceedings of the National Academy of Sciences, 110(13), 1181-1190.

Freedman, D., Pisani, R., Purves, R., (1998) *Statistics*. New York:Norton.

Lindstrom, M. (2016) *Small Data: The Tiny Clues That Uncover Huge Trends*. London:St. Martin's Press.

Meng, X.L. (2014). A Trio of Inference Problems that Could Win You a Nobel Prize in Statistics (If You Help Fund It). In Past, Present, and Future of Statistical Science (Eds: X. Lin, et. al). Boca Raton:CRC Press

Secchi, P. (2018). On the role of statistics in the era of big data: a call for debate. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear*