Statistical Science
2005, Vol. 20, No. 3, 231–238
DOI 10.1214/088342305000000269

How to Lie with Bad Data

Richard D. De Veaux and David J. Hand

Abstract. As Huff's landmark book made clear, lying with statistics can be accomplished in many ways. Distorting graphics, manipulating data or using biased samples are just a few of the tried and true methods. Failing to use the correct statistical procedure or failing to check the conditions for when the selected method is appropriate can distort results as well, whether the motives of the analyst are honorable or not. Even when the statistical procedure and motives are correct, bad data can produce results that have no validity at all. This article provides some examples of how bad data can arise, what kinds of bad data exist, how to detect and measure bad data, and how to improve the quality of data that have already been collected.

Key words and phrases: Data quality, data profiling, data rectification, data consistency, accuracy, distortion, missing values, record linkage, data warehousing, data mining.

1. INTRODUCTION

Bad data can ruin any analysis. "Garbage in, garbage out" is as true today in this era of terabytes of data and distributed computing as it was in 1954 when *How to Lie with Statistics* was published (Huff, 1954). Distortions in the data are likely to produce distortions in the conclusions, to the extent that these may be wildly inaccurate, completely invalid or useless.

The cost of bad data can be enormous. Estimates of how much bad data cost U.S. industry permeate industry publications and the Internet. PricewaterhouseCoopers (2004), in a recent survey of "Top 500" corporations, found that most corporations are experiencing major impacts to their business as a result of poor data quality. In their survey, 75% of respondents reported significant problems as the result of defective data. David Loshin, author of *Enterprise Knowledge Management: The Data Quality Approach*, states that "scrap and rework attributable to poor data quality accounts for 20–25% of an organization's budget" (Loshin, 2001). An A. T. Kearney study of the retail consumer products supply chain concluded that "bad data costs the electro industry $1.2B annually." While the accuracy of these claims is hard to verify, it is clear that data quality is a concern to business worldwide. An informal survey of topics in management seminars shows the prevalence of data quality as an important topic and concern for top-level executives and managers.

Anyone who has analyzed real data knows that the majority of their time on a data analysis project will be spent "cleaning" the data before doing any analysis. Common wisdom puts the extent of this at 60–95% of the total project effort, and some studies (Klein, 1998) suggest that "between one and ten percent of data items in critical organizational databases are estimated to be inaccurate" (Laudon, 1986; Madnick and Wang, 1992; Morey, 1982; Redman, 1992). Somewhat paradoxically, most statistical training assumes that the data arrive "precleaned." Students, whether in Ph.D. programs or in an undergraduate introductory course, are not routinely taught how to check data for accuracy or even to worry about it. Exacerbating the problem further are claims by software vendors that their techniques can produce valid results no matter what the quality of the incoming data.

How pervasive are bad data? Not only is industry concerned with bad data, but examples are ubiquitous in the scientific literature of the past 50 years as well. In the 1978 Fisher Lecture "Statistics in Society: Problems Unsolved and Unformulated (Kruskal,

Richard D. De Veaux is Professor, Department of Mathematics and Statistics, Bronfman Science Center, Williams College, Williamstown, Massachusetts 01267, USA (e-mail: deveaux@williams.edu). David J. Hand is Professor, Department of Mathematics, Imperial College, London SW7 2AZ, UK (e-mail: d.j.hand@imperial.ac.uk).

1981), Kruskal devoted much of his time to "inconsistent or clearly wrong data, especially in large data sets." As just one example, he cited a 1960 census study that showed 62 women, aged 15 to 19 with 12 or more children. Coale and Stephan (1962) pointed out similar anomalies when they found a large number of 14-year-old widows. In a classic study by Wolins (1962), a researcher attempted to obtain raw data from 37 authors of articles appearing in American Psychological Association journals. Of the seven data sets that were actually obtained, three contained gross data errors.

A 1986 study by the U.S. Census estimated that between 3 and 5% of all census enumerators engaged in some form of fabrication of questionnaire responses without actually visiting the residence. This practice was widespread enough to warrant its own term: curbstoning, which is the "enumerator jargon for sitting on the curbstone filling out the forms with made-up information" (Wainer, 2004). While curbstoning does not imply bad data per se, at the very least, such practices imply that the data set you are analyzing does not describe the underlying mechanism you think you are describing.

What exactly are bad data? The quality of data is relative both to the context and to the question one is trying to answer. If data are wrong, then they are obviously bad, but context can make the distinction more subtle. In a regression analysis, errors in the predictor variables may bias the estimates of the regression coefficients and this will matter if the aim hinges on interpreting these values, but it will not matter if the aim is predicting response values for new cases drawn from the same distribution. Likewise, whether data are "good" also depends on the aims: precise, accurate measurements are useless if one is measuring the wrong thing. Increasingly in the modern world, especially in data mining, we are confronted with secondary data analysis: the analysis of data that have been collected for some other purpose (e.g., analyzing billing data for transaction patterns). The data may have been perfect for the original aim, but could have serious deficiencies for the new analysis.

For this paper, we will take a rather narrow view of data quality. In particular, we are concerned with data accuracy, so that, for us, "poor quality data are defined as erroneous values assigned to attributes of some entity," as in Pierce (1997). A broader perspective might also take account of relevance, timeliness, existence, coherence, completeness, accessibility, security and other data attributes. For many problems, for example, data gradually become less and less relevant—a phenomenon sometimes termed data decay or population drift (Hand, 2004a). Thus the characteristics collected on mortgage applicants 25 years ago would probably not be of much use for developing a predictive risk model for new applicants, no matter how accurately they were measured at the time. In some environments, the time scale that renders a model useless can become frighteningly short. A model of customer behavior on a web site may quickly become out of date. Sometimes different aspects of this broader interpretation of data quality work in opposition. Timeliness and accuracy provide an obvious example (and, indeed, one which is often seen when economic time series are revised as more accurate information becomes available).

From the perspective of the statistical analyst, there are three phases in data evolution: collection, preliminary analysis and modeling. Of course, the easiest way to deal with bad data is to prevent poor data from being collected in the first place. Much of sample survey methodology and experimental design is devoted to this subject, and many famous stories of analysis gone wrong are based on faulty survey designs or experiments. The *Literary Digest* poll proclaiming Landon's win over Roosevelt in 1936 that starred in Chapter 1 of Huff (1954) is just one of the more famous examples. At the other end of the process, we have resistant and robust statistical procedures explicitly designed to perform adequately even when a percentage of the data do not conform or are inaccurate, or when the assumptions of the underlying model are violated.

In this article we will concentrate on the "middle" phase of bad data evolution—that is, on its discovery and correction. Of course, no analysis proceeds linearly through the process of initial collection to final report. The discoveries in one phase can impact the entire analysis. Our purpose will be to discuss how to recognize and discover these bad data using a variety of examples, and to discuss their impact on subsequent statistical analysis. In the next section we discuss the causes of bad data. Section 3 discusses the ways in which data can be bad. In Section 4 we turn to the problem of detecting bad data and in Section 5 we provide some guidelines for improving data quality. We summarize and present our conclusions in Section 6.

2. WHAT ARE THE CAUSES OF BAD DATA?

There is an infinite variety to the ways in which data can go bad, and the specifics depend on the underlying

process that generate the data. Data may be distorted from the outset during the initial collection phase or they may be distorted when the data are transcribed, transferred, merged or copied. Finally, they may deteriorate, change definition or otherwise go through transformations that render them less representative of the original underlying process they were designed to measure.

The breakdown in the collection phase can occur whether the data are collected by instrument or directly recorded by human beings. Examples of breakdowns at the instrument level include instrument drift, initial miscalibration, or a large random or otherwise unpredictable variation in measurement. As an example of instrument level data collection, consider the measurement of the concentration of a particular chemical compound by gas chromatography, as used in routine drug testing. When reading the results of such a test, it is easy to think that a machine measures the amount of the compound in an automatic and straightforward way, and thus that the resulting data are measuring some quantity directly. It turns out to be a bit more complicated. At the outset, a sample of the material of interest is injected into a stream of carrier gas where it travels down a silica column heated by an oven. The column then separates the mixture of compounds according to their relative attraction to a material called the adsorbent. This stream of different compounds travels "far enough" (via choices of column length and gas flow rates) so that by the time they pass by the detector, they are well separated (at least in theory). At this point, both the arrival time and the concentration of the compound are recorded by an electromechanical device (depending on the type of detector used). The drifts inherent in the oven temperature, gas flow, detector sensitivity and a myriad of other environmental conditions can affect the recorded numbers. To determine actual amounts of material present, a known quantity must be tested at about the same time and the machine must be calibrated. Thus the number reported as a simple percentage of compound present has not only been subjected to many potential sources of error in its raw form, but is actually the output of a calibration model.

Examples of data distortion at the human level include misreading of a scale, incorrect copying of values from an instrument, transposition of digits and misplaced decimal points. Of course, such mistakes are not always easy to detect. Even if every data value is checked for plausibility, it often takes expert knowledge to know if a data value is reasonable or absurd. Consider the report in *The Times* of London that some surviving examples of the greater mouse-eared bat, previously thought to be extinct, had been discovered hibernating in West Sussex. It went on to assert that "they can weigh up to 30 kg" (see Hand, 2004b, Chapter 4). A considerable amount of entertaining correspondence resulted from the fact that they had misstated the weight by three decimal places.

Sometimes data are distorted from the source itself, either knowingly or not. Examples occur in survey work and tax returns, just to name two. It is well known to researchers of sexual behavior that men tend to report more lifetime sexual partners than women, a situation that is highly unlikely sociologically (National Statistics website: www.statistics.gov.uk). Some data are deliberately distorted to prevent disclosure of confidential information collected by governments in, for example, censuses (e.g., Willenborg and de Waal, 2001) and health care data.

Even if the data are initially recorded accurately, data can be compromised by data integration, data warehousing and record linkage. Often a wide range of sources of different types are involved (e.g., in the pharmaceutical sector, data from clinical trials, animal trials, manufacturers, marketing, insurance claims and postmarketing surveillance might be merged). At a more mundane level, records that describe different individuals might be inappropriately merged because they are described by the same key. When going through his medical records for insurance purposes, one of the authors discovered that he was recorded as having had his tonsils removed as a child. A subsequent search revealed the fact that the records of someone else with the same name (but a different address) had been mixed in with his. More generally, what is good quality for (the limited demands made of) an operational data base may not be good quality for (potentially unlimited demands made of) a data warehouse.

In a data warehouse, the definitions, sources and other information for the variables are contained in a dictionary, often referred to as metadata. In a large corporation it is often the IT (information technology) group that has responsibility for maintaining both the data warehouse and metadata. Merging sources and checking for consistent definitions form a large part of their duties.

A recent example in bioinformatics shows that data problems are not limited to business and economics. In a recent issue of *The Lancet*, Petricoin et al. (2002) reported an ability to distinguish between serum samples from healthy women, those with ovarian cancers and women with a benign ovarian disease. It was so

exciting that it prompted the "U.S. Congress to pass a resolution urging continued funding to drive a new diagnostic test toward the clinic" (Check, 2004). The researchers trained an algorithm on 50 cancer spectra and 50 normals, and then predicted 116 new spectra. The results were impressive with the algorithm correctly identifying all 50 of the cancers, 47 out of 50 normals, and classifying the 16 benign disease spectra as "other." Statisticians Baggerly, Morris and Coombes (2004) attempted to reproduce the Petricoin et al. results, but were unable to do so. Finally, they concluded that the three types of spectra had been preprocessed differently, so that the algorithm correctly identified differences in the data, much of which that had nothing to do with the underlying biology of cancer.

A more subtle source of data distortion is a change in the measurement or collection procedure. When the cause of the change is explicit and recognized, this can be adjusted for, at least to some extent. Common examples include a change in the structure of the Dow Jones Industrial Average or the recent U.K. change from the Retail Price Index to the European Union standard Harmonized Index of Consumer Prices. In other cases, one might not be aware of the change. Some of the changes can be subtle. In looking at historical records to assess long-term temperature changes, Jones and Wigley (1990) noted "changing landscapes affect temperature readings in ways that may produce spurious temperature trends." In particular, the location of the weather station assigned to a city may have changed. During the 19th century, most cities and towns were too small to impact temperature readings. As urbanization increased, urban heat islands directly affected temperature readings, creating bias in the regional trends. While global warming may be a contributor, the dominant factor is the placement of the weather station, which moved several times. As it became more and more surrounded by the city, the temperature increased, mainly because the environment itself had changed.

A problem related to changes in the collection procedure is not knowing the true source of the data. In scientific analysis, data are often preprocessed by technicians and scientists before being analyzed. The statistician may be unaware (or uninterested) in the details of the processing. To create accurate models, however, it can be important to know the source and therefore the accuracy of the measurements. Consider a study of the effect of ocean bottom topography on sea ice formation in the southern oceans (De Veaux, Gordon, Comiso and Bacherer, 1993). After learning that wind can have a strong effect on sea ice formation, the statistician, wanting to incorporate this predictor into a model, asked one of the physicists whether any wind data existed. It was difficult to imagine very many Antarctic stations with anemometers and so he was very surprised when the physicist replied, "Sure, there's plenty of it." Excitedly he asked what spatial resolution he could provide. When the physicist countered with "what resolution do you want?" the statistician became suspicious. He probed further and asked if they really had anemometers set up on a 5 km grid on the sea ice? He said, "Of course not. The wind data come from a global weather model—I can generate them at any resolution you want!" It turned out that *all* the other satellite data had gone through some sort of preprocessing before it was given to the statistician. Some were processed actual direct measurements, some were processed through models and some, like the wind, were produced *solely* from models. Of course, this (as with curbstoning) does not necessarily imply that the resulting data are bad, but it should at least serve to warn the analyst that the data may not be what they were thought to be.

Each of these different mechanisms for data distortions has its own set of detection and correction challenges. Ensuring good data collection through survey and/or experimental design is certainly an important first step. A bad design that results in data that are not representative of the phenomenon being studied can render even the best analysis worthless. At the next step, detecting errors can be attempted in a variety of ways, a topic to which we will return in Section 4.

3. IN HOW MANY WAYS?

Data can be bad in an infinite variety of ways, and some authors have attempted to construct taxonomies of data distortion (e.g., Kim et al., 2003). An important simple categorization is into missing data and distorted values.

3.1 Missing Data

Data can be missing at two levels: entire records might be absent, or one or more individual fields may be missing. If entire records are missing, any analysis may well be describing or making inferences about a population different from that intended. The possibility that entire records may be missing is particularly problematic, since there will often be no way of knowing this. Individual fields can be missing for a huge variety of reasons, and the mechanism by which they are missing is likely to influence their distribution over the

data, but at least when individual fields are missing one can see that this is the case.

If the missingness of a particular value is unrelated either to the response or predictor variables (missing completely at random—Little and Rubin, 1987, give technical definitions), then case deletion can be employed. However, even ignoring the potential bias problems, complete case deletion can severely reduce the effective sample size. In many data mining situations with a large number of variables, even though each field has only a relatively small proportion of missing values, all of the records may have some values missing, so that the case deletion strategy leaves one with no data at all.

Complications arise when the pattern of missing data does depend on the values that would have been recorded. If, for example, there are no records for patients who experience severe pain, inferences to the entire pain distribution will be impossible (at least, without making some pretty strong distributional assumptions). Likewise, poor choice of a missing value code (e.g., 0 or 99 for age) or accidental inclusion of a missing value code in the analysis (e.g., 99,999 for age) has been known to lead to mistaken conclusions.

Sometimes missingness arises because of the nature of the problem, and presents real theoretical and practical issues. For example, in personal banking, banks accept those loan applicants whom they expect to repay the loans. For such people, the bank eventually discovers the true outcome (repay, do not repay), but for those rejected for a loan, the true outcome is unknown: it is a missing value. This poses difficulties when the bank wants to construct new predictive models (Hand and Henley, 1993; Hand, 2001). If a loan application asks for household income, replacing a missing value by a mean or even by a model based imputation may lead to a highly optimistic assessment of risk.

When the missingness in a predictor is related directly to the response, it may be useful for exploratory and prediction purposes to create indicator variables for each predictor, where the variable is a binary indictor of whether the variable is missing or not. For categorical predictor variables, missing values can be treated simply as a new category. In a study of dropout rates from a clinical trial for a depression drug, it was found that the single most important indicator of ultimately dropping out from the study was not the depression score on the second week's test as indicated from complete case analysis, but simply the indicator of whether the patient showed up to take it (De Veaux, Donahue and Small, 2002).

3.2 Distorted Data

Although there are an unlimited number of possible causes of distortion, a first split can be made into those attributable to instrumentation and those attributed to human agency. Floor and ceiling effects are examples of the first kind (instruments here can be mechanical or electronic, but also questionnaires), although in this case it is sometimes possible to foresee that such things might occur and take account of this in the statistical modeling. Human distortions can arise from misreading instruments or misrecording values at any level. Brunskill (1990) gave an illustration from public records of birth weights, where ounces are commonly confused with pounds, the number 1 is confused with 11 and errors in decimal placements produce order of magnitude errors. In such cases, using ancillary information such as gestation times or newborn heights can help to spot gross errors. Some data collection procedures, in an attempt to avoid missing data, actually introduce distortions. A data set we analyzed had a striking number of doctors born on November 11, 1911. It turned out that most doctors (or their secretaries) wanted to avoid typing in age information, but because the program insisted on a value and the choice of 00/00/00 was invalid, the easiest way to bypass the system was simply to type 11/11/11. Such errors might not seem of much consequence, but they can be crucial. Confusion between English and metric units was responsible for the loss of the $125 million Martian Climate Orbiter space probe (*The New York Times*, October 1, 1999). Jet Propulsion Laboratory engineers mistook acceleration readings measured in English units of pound-seconds for the metric measure of force in newton-seconds. In 1985, in a precedence setting case, the Supreme Court ruled that Dun & Bradstreet had to pay $350,000 in libel damages to a small Vermont construction company. A part-time student worker had apparently entered the wrong data into the Dun & Bradstreet data base. As a result, Dun & Bradstreet issued a credit report that mistakenly identified the construction company as bankrupt (Percy, 1986).

4. HOW TO DETECT DATA ERRORS

While it may be obvious that a value is missing from a record, it is often less obvious that a value is in error. The presence of errors can (sometimes) be proven, but the absence of errors cannot. There is no guarantee that a data set that looks perfect will not contain mistakes. Some of these mistakes may be intrinsically

undetectable: they might be values that are well within the range of the data and could easily have occurred. Moreover, since errors can occur in an unlimited number of ways, there is no end to the list of possible tests for detecting errors. On the other hand, strategic choice of tests can help to pinpoint the root causes that lead to errors and, hence, to the identification of changes in the data collection process that will lead to the greatest improvement in data quality.

When the data collection can be repeated, the results of the duplicate measurements, recordings or transcriptions (e.g., the double entry system used in clinical trials) can be compared by automatic methods. In this "duplicate performance method," a machine checks for any differences in the two data records. All discrepancies are noted and the only remaining errors are when *both* collectors made the same mistake. Strayhorn (1990) and West and Winkler (1991) provided statistical methods for estimating that proportion. In another quality control method, known errors are added to a data set whose integrity is then assessed by an external observer. The "known errors" method devised statistical methods for estimating how many errors remain based on the success of the observer in discovering the known errors (Strayhorn, 1990; West and Winkler, 1991). Taking this further, one can build models (similar to those developed for software reliability) that estimate how many errors are likely to remain in a data set based on extrapolation from the rate of discovery of errors. At some point one decides that the impact of remaining errors on the conclusions is likely to be sufficiently small that one can ignore them.

Automatic methods of data collection use metadata information to check for consistency across multiple records or variables, integrity (e.g., correct data type), plausibility (within the possible range of the data) and coherence between related variables (e.g., number of sons plus number of daughters equals number of children). Sometimes redundant data can be collected with such checks in mind. However, one cannot rely on software to protect one from mistakes. Even when such automatic methods are in place, the analyst should spend some time looking for errors in the data prior to any modeling effort.

Data profiling is the use of exploratory and data mining tools aimed at identifying errors, rather than at the substantive questions of interest. When the number of predictor variables is manageable, simple plots such as bar charts, histograms, scatterplots and time series plots can be invaluable. The human eye has evolved to detect anomalies, and this should be taken advantage of by presenting the data in a form whereby advantage can be taken of these abilities. Such plots have become prevalent in statistical packages for examining missing data patterns. Hand, Blunt, Kelly and Adams (2000) gave the illustration of a plot showing a point for each missing value in a rectangular array of 1012 potential sufferers from osteoporosis measured on 45 variables. It is immediately clear which cases and which variables account for most of the problems.

Unfortunately, as we face larger and larger data sets, so we are also faced with increasing difficulty in data profiling. The missing value plot described above works for a thousand cases, but would probably not be so effective for 10 million. Even in this case, however, a Pareto chart of percent missing for each variable may be useful for deciding where to spend data preparation effort. Knowing that a variable is 96% missing makes one think pretty hard about including it in a model. On the other hand, separate manual examination of each of 30,000 gene expression variables is not to be recommended.

When even simple summaries of all the variables in a data base are not feasible, some methods for reducing the number of potential predictors in the models might be warranted. We see an important role for data mining tools here. It may be wise to reverse the usual paradigm of explore the data first, then model. Instead, exploratory models of the data can be useful as a *first step* and can serve two purposes (De Veaux, 2002). First, models such as tree models and clustering can highlight groups of anomalous cases. Second, the models can be used to reduce the number of potential predictor variables and enable the analyst to examine the remaining predictors in more detail. The resulting process is a circular one, with more examination possible at each subsequent modeling phase. Simply checking whether 500 numerical predictor variables are categorical or quantitative without the aid of metadata is a daunting (and tedious) task. In one analysis, we were asked to develop a fraud detection model for a large credit card bank. In the data set was one potential predictor variable that ranged from around 2000 to 9000, roughly symmetric and unimodal, which was selected as a highly significant predictor for fraud in a stepwise logistic regression model. It turned out that this predictor was a categorical variable (SIC code) used to specify the industry from which the product purchases in the transaction came. Useless as a predictor in a logistic regression model, it had escaped detection as a categorical variable among the several hundred potential candidates. Once the preliminary model whittled the

candidate predictors down to a few dozen, it was easy to use standard data analysis techniques and to detect which were appropriate for the final model.

5. IMPROVING DATA QUALITY

The best way to improve the quality of data is to improve things in the data collection phase. The ideal would be to prevent errors from arising in the first place. Prevention and detection have a reciprocal role to play here. Once one has detected data errors, one can investigate why they occurred and prevent them from happening in the future. Once it has been recognized (detected) that the question "How many miles do you commute to work each day?" permits more than one interpretation, mistakes can be prevented by rewording. Progress toward direct keyboard or other electronic data entry systems means that error detection tools can be applied in real time at data entry—when there is still an opportunity to correct the data. At the data base phase, metadata can be used to ensure that the data conform to expected forms, and relationships between variables can be used to cross-check entries. If the data can be collected more than once, the rate of discovery of errors can be used as the basis for a statistical model to reveal how many undetected errors are likely to remain in the data base.

Various other principles also come into play when considering how to improve data quality. For example, a Pareto principle often applies: most of the errors are attributable to just a few variables. This may happen simply because some variables are intrinsically less reliable (and important) than others. Sometimes it is possible to improve the overall level of quality significantly by removing just a few of these low quality variables. This has a complementary corollary: a law of diminishing returns applies that suggests that successive attempts to improve the quality of the data are likely to lead to less improvement. If one has a particular analytic aim in mind, then one might reasonably assert that data errors that do not affect the conclusions do not matter. Moreover, for those that do matter, perhaps the ease with which they can be corrected should have some bearing on the effort that goes into detecting them—although the overriding criterion should be the loss consequent on the error being made. This is allied with the point that the base rate of errors should be taken into account: if one expects to find many errors, then it is worth attempting to find them, since the likely rewards, in terms of an improved data base, are likely to be large. In a well-understood environment, it might even be possible to devise useful error detection and correction resource allocation strategies.

Sometimes an entirely different approach to improving data quality can be used. This is simply to hide the poor quality by coarsening or aggregating the data. In fact, a simple example of this implicitly occurs all the time: rather than reporting uncertain and error-prone final digits of measured variables, researchers round to the nearest digit.

6. CONCLUSIONS AND FURTHER DISCUSSION

This article has been about data quality from the perspective of an analyst called upon to extract some meaning from it. We have already remarked that there are also other aspects to data quality, and these are of equal importance when action is to be taken or decisions made on the basis of the data. These include such aspects as timeliness (the most sophisticated analysis applied to out-of-date data will be of limited value), completeness and, of central importance, fitness for purpose. Data quality, in the abstract, is all very well, but what may be perfectly fine for one use may be woefully inadequate for another. Thus ISO 8402 defines quality as "The totality of characteristics of an entity that bare on its ability to satisfy stated and implied needs."

It is also important to maintain a sense of proportion in assessing and deciding how to cope with data distortions. In one large quality control problem in polymer viscosity, each 1% improvement was worth about $1,000,000 a year, but viscosity itself could be measured only to a standard deviation of around 8%. Before bothering about the accuracy of the predictor variables, it was first necessary to find improved ways to measure the response. In an entirely different context, much work in the personal banking sector concentrates on improved models for predicting risk—where, again, a slight improvement translates into millions of dollars of increased profit. In general, however, these models are based on retrospective data—data drawn from distributions that are unlikely still to apply. We need to be sure that the inaccuracies induced by this population drift do not swamp the apparent improvements we have made.

Data quality is a key issue throughout science, commerce, and industry, and entire disciplines have grown up to address particular aspects of the problem. In manufacturing and, to a lesser extent, the service industries, we have schools for quality control and total quality management (Six Sigma, Kaizen, etc.). In large part,

these are concerned with reducing random variation. In official statistics, strict data collection protocols are typically used.

Of course, ensuring high quality data does not come without a cost. The bottom line is that one must weigh up the potential gains to be made from capturing and recording better quality data against the costs of ensuring that quality. No matter how much money one spends, and how much resource one consumes in attempting to detect and prevent bad data, the unfortunate fact is that bad data will always be with us.

REFERENCES

BAGGERLY, K. A, MORRIS, J. S. and COOMBES, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different experiments. *Bioinformatics* **20** 777–785.

BRUNSKILL, A. J. (1990). Some sources of error in the coding of birth weight. *American J. Public Health* **80** 72–73.

CHECK, E. (2004). Proteomics and cancer: Running before we can walk? *Nature* **429** 496–497.

COALE, A. J. and STEPHAN, F. F. (1962). The case of the Indians and the teen-age widows. *J. Amer. Statist. Assoc.* **57** 338–347.

DE VEAUX, R. D. (2002). Data mining: A view from down in the pit. *Stats* (34) 3–9.

DE VEAUX, R. D., DONAHUE, R. and SMALL, R. D. (2002). Using data mining techniques to harvest information in clinical trials. Presentation at Joint Statistical Meetings, New York.

DE VEAUX, R. D., GORDON, A., COMISO, J. and BACHERER, N. E. (1993). Modeling of topographic effects on Antarctic sea-ice using multivariate adaptive regression splines. *J. Geophysical Research—Oceans* **98** 20,307–20,320.

HAND, D. J. (2001). Reject inference in credit operations. In *Handbook of Credit Scoring* (E. Mays, ed.) 225–240. Glenlake Publishing, Chicago.

HAND, D. J. (2004a). Academic obsessions and classification realities: Ignoring practicalities in supervised classification. In *Classification, Clustering and Data Mining Applications* (D. Banks, L. House, F. R. McMorris, P. Arabie and W. Gaul, eds.) 209–232. Springer, Berlin.

HAND, D. J. (2004b). *Measurement Theory and Practice*: *The World Through Quantification*. Arnold, London.

HAND, D. J., BLUNT, G., KELLY, M. G. and ADAMS, N. M. (2000). Data mining for fun and profit (with discussion). *Statist. Sci.* **15** 111–131.

HAND, D. J. and HENLEY, W. E. (1993). Can reject inference ever work? *IMA J. of Mathematics Applied in Business and Industry* **5**(4) 45–55.

HUFF, D. (1954). *How to Lie with Statistics*. Norton, New York.

JONES, P. D. and WIGLEY, T. M. L. (1990). Global warming trends. *Scientific American* **263**(2) 84–91.

KIM, W., CHOI, B.-J., HONG, E.-K., KIM, S.-K. and LEE, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery* **7** 81–99.

KLEIN, B. D. (1998). Data quality in the practice of consumer product management: Evidence from the field. *Data Quality* **4**(1).

KRUSKAL, W. (1981). Statistics in society: Problems unsolved and unformulated. *J. Amer. Statist. Assoc.* **76** 505–515.

LAUDON, K. C. (1986). Data quality and due process in large interorganizational record systems. *Communications of the ACM* **29** 4–11.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

LOSHIN, D. (2001). *Enterprise Knowledge Management*: *The Data Quality Approach*. Morgan Kaufmann, San Francisco.

MADNICK, S. E. and WANG, R. Y. (1992). Introduction to the TDQM research program. Working Paper 92-01, Total Data Quality Management Research Program.

MOREY, R. C. (1982). Estimating and improving the quality of information in a MIS. *Communications of the ACM* **25** 337–342.

PERCY, T. (1986). My data, right or wrong. *Datamation* **32**(11) 123–124.

PETRICOIN, E. F., III, ARDEKANI, A. M., HITT, B. A., LEVINE, P. J., FUSARO, V. A., STEINBERG, S. M., MILLS, G. B., SIMONE, C., FISHMAN, D. A., KOHN, E. C. and LIOTTA, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* **359** 572–577.

PIERCE, E. (1997). Modeling database error rates. *Data Quality* **3**(1). Available at www.dataquality.com/dqsep97.htm.

PRICEWATERHOUSECOOPERS (2004). *The Tech Spotlight* **22**. Available at www.pwc.com/extweb/manissue.nsf/docid/2D6E2F57E06E022F85256B8F006F389A.

REDMAN, T. C. (1992). *Data Quality. Management and Technology*. Bantam, New York.

STRAYHORN, J. M. (1990). Estimating the errors remaining in a data set: Techniques for quality control. *Amer. Statist.* **44** 14–18.

WAINER, H. (2004). Curbstoning IQ and the 2000 presidential election. *Chance* **17**(4) 43–46.

WEST, M. and WINKLER, R. L. (1991). Data base error trapping and prediction. *J. Amer. Statist. Assoc.* **86** 987–996.

WILLENBORG, L. and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. Springer, New York.

WOLINS, L. (1962). Responsibility for raw data. *American Psychologist* **17** 657–658.