Statistical Science
2000, Vol. 15, No. 2, 111–131

Data Mining for Fun and Profit

David J. Hand, Gordon Blunt, Mark G. Kelly and Niall M. Adams

Abstract. Data mining is defined as the process of seeking interesting or valuable information within large data sets. This presents novel challenges and problems, distinct from those typically arising in the allied areas of statistics, machine learning, pattern recognition or database science. A distinction is drawn between the two data mining activities of model building and pattern detection. Even though statisticians are familiar with the former, the large data sets involved in data mining mean that novel problems do arise. The second of the activities, pattern detection, presents entirely new classes of challenges, some arising, again, as a consequence of the large sizes of the data sets. Data quality is a particularly troublesome issue in data mining applications, and this is examined. The discussion is illustrated with a variety of real examples.

Key words and phrases: Data mining, knowledge discovery, large data sets, computers, databases.

1. INTRODUCTION

Data mining is the process of seeking interesting or valuable information within large databases. The novelty, and the reason that a new term has been coined to describe the activity, has its origin primarily in the large sizes of modern databases. The size means that standard statistical exploratory data analysis procedures need to be extended, modified, adapted and also supplemented by different kinds of procedures. The result is that data mining is an interdisciplinary subject, representing the confluence of ideas from statistics, exploratory data analysis, machine learning, pattern recognition, database technology, and other disciplines.

The size of modern databases is illustrated by the following examples. Barclaycard, the U.K.'s largest credit card company, carries out 350 million transactions a year. However, this is nothing compared to the American retailer Wal-Mart, which makes over 7 billion transactions a year (Babcock, 1994). More strikingly still, according to Cortes and Pregibon (1997), AT&T carries over 70 billion long distance calls annually. Harrison (1993) remarks that Mobil Oil aims to store over 100 terabytes of data concerned with oil exploration. Fayyad, Piatetsky-Shapiro, and Smyth (1996) say that the NASA Earth Observing System was projected to generate on the order of 50 gigabytes of data per hour around the turn of the century. The human genome project has already collected gigabytes of data.

Numbers as large as those above are very much a consequence of modern electronics, computer and database technology. The data may have been collected as secondary to some other activity, or to answer a specific question but then retained. Once collected, they clearly represent some kind of resource: it is likely, arguably certain, that such mountains of data contain valuable or interesting information, if only one could identify and extract it.

The term "data mining" is not a new one to statisticians. However, rather than the promise implicit in the above (the information is there in the data, and "all we have to do" is extract it), it often carries negative connotations because one can *always* find *apparent* structures in data sets. Many of these structures will not be real in the sense that they represent aspects of an underlying distribution, but will be attributable to the random aspects of

David J. Hand is Professor, Department of Mathematics, Imperial College of Science, Technology and Medicine, Huxley Building, 180 Queen's Gate, London SW7 2BZ, U.K. Gordon Blunt is Marketing Analyst, Marketing Department, Barclaycard, 1234 Pavilion Drive, Northampton, NN4 7SG, U.K. Mark G. Kelly is Research Fellow, Department of Mathematics, Imperial College, Huxley Building, 180 Queen's Gate, London SW7 2BZ, U.K. Niall M. Adams is Research Fellow, Department of Mathematics, Imperial College, Huxley Building, 180 Queen's Gate, London SW7 2BZ, U.K.

the data generating process. Statisticians, to whom *inference* is a fundamental activity, are acutely aware of this and have as a central concern the question of how to distinguish between the underlying "systematic" components and the random components of data. Thus, on observing a small local cluster of data points, one of the statistician's chief concerns may be whether the clustering could reasonably be attributable to chance or not. In contrast, data mining practitioners concern themselves primarily with identifying potentially interesting or valuable structures in data (i.e., with *finding* the "small local cluster of data points" in the first place), shifting the responsibility for determining "reality" to the database owner or domain expert. This also allows the expert to characterize structures which, though real, are not of interest (for example, a data mining exercise may reveal that almost all sufferers from prostate cancer in the database are male). We have more to say about such issues below.

Chance is one source of structures in data which have no matching underlying "reality," or which are not valuable or interesting. Another source is data corruption. One should expect any large data set to be imperfect. This poses particular challenges for those concerned with finding interesting, valuable and meaningful structures in large data sets. It means that one may discover structures which are an aspect of systematic variation between objects, but which have arisen because of missing data, distorted data or ambiguity of definition (e.g., a few sufferers from prostate cancer in the illustration mentioned above who are apparently female). Such structures may be of interest, but equally may not be.

We commented above that most statisticians are concerned with inference of one kind or another. Their aim is to take a sample of data and make a statement, with associated probabilities, about the population from which it came. This may be at a low level—a simple hypothesis test—or at a higher level—an overall model. Some data mining tasks are of this kind: in some situations, as we describe in Section 2.1, the analysis can be based on a sample from a large database. Others, however, especially *pattern detection* problems (see below), cannot be based on a sample. Moreover, in some cases one has the entire data set available (in the numerical examples above, *all* of the Wal-Mart transactions for a year will be available, not merely a sample of them). Again this can lead to differences from standard statistical approaches.

Because data mining is concerned with discovering structure within existing databases, it is seldom concerned with issues of data collection. In particular, sciences of efficient data collection, such as experimental design, survey design and questionnaire design, are beyond its remit. This is one way in which data mining differs from statistics. Another way is that data mining places greater emphasis on algorithms than does statistics. That data mining should be more concerned with algorithms is hardly surprising. Given the large sizes of the data sets that may be examined, one must rely very heavily on automatic data processing of some kind. There is no way that one can individually examine a billion data points. Moreover, the algorithms have to be fast. This has led to an interest in adaptive and sequential estimation methods, in which estimates are updated data point by data point, so minimizing the number of disc accesses. Machine learning, as the very word "learning" suggests, has historically placed emphasis on such approaches. Perhaps because of this, in contrast to most modern statistics, where the model is central and deriving a model is the goal of statistical analysis (and from which other questions can be answered), to many data mining practitioners the algorithm is central. Often they may not even think in terms of a model-building process at all, instead viewing it as a data driven descriptive exercise, with the algorithm determining what sort of description emerges. Hand (1999) has pointed out that the Gifi school of multivariate statistics also adopts this perspective, so it is not without precedent, even amongst statisticians. (Gifi 1990, page 34) says: "In this book we adopt the point of view that, given some of the most common MVA [multivariate analysis] questions, it is possible to start either from the model or from the technique. As we have seen in Section 1.1 classical multivariate statistical analysis starts from the model, ... In many cases, however, the choice of the model is not at all obvious, choice of a conventional model is impossible, and computing optimum procedures is not feasible. In order to do something reasonable in these cases we start from the other end, with a class of techniques designed to answer the MVA questions, and postpone the choice of model and of optimality criterion."

In the above, we have described the objectives of data mining as being to find structure in data. It is useful to distinguish between two types of structure (Hand, 1998a): *models* and *patterns*. Both are widely sought in data mining exercises. The distinction, though sometimes blurred, is nevertheless a useful one. A *model* is an overall summary of a set of data, or a subset of the data. This is thus the standard statistical usage. We can speak of a regression model, a Box-Jenkins time series model, a dynamic linear model, a conditional independence graph model, a cluster model and so on. Such structures represent a large-scale summary of a mass of data.

In contrast, a *pattern* is a local structure, possibly (though not necessarily) referring to only a relatively small number of objects. ("Small" here could mean 10 or 100,000. It depends on the context and, of course, the size of the overall data set.) Thus a pattern might be an anomalously high log-odds ratio for a small group of objects or a small sequence of values which occurs several times in a time series trace (this is precisely the sort of thing that chartists seek in stock prices). A multiway table of counts can be analysed via a log-linear model, which will summarize its broad features, or one can try to identify unexpectedly high or low cell counts. This suggests that patterns are defined relative to a model. This is often the case—outliers are another example—but it is not always so; as we illustrate below, the pattern may simply be defined relative to a broad notion of continuity, something which is really too general to call a "model."

As with data analysis in general, data mining is not a once-off activity. That is, presented with a billion point data set, one does not simply mine it and be done with it. Rather, the exercise should be seen as an interactive *process* involving both the data miner and the domain expert, as well as the data.

In the next section we examine the sorts of tools being used and developed for data mining applications. We expand on the distinction between models and patterns, and illustrate with some real data sets. Section 3 discusses some of the challenging problems which arise as a consequence of the sheer size of the data sets which are becoming available, and Section 4 looks at the vital issue of data quality. Section 5 draws some overall conclusions.

2. DATA MINING TOOLS

In Section 1 we distinguished between models and patterns. Any kind of statistical model may appear in a data mining application, and we need not dwell on tools for these here, since they will be familiar to statisticians. Examination of recent data mining conference proceedings shows that certain classes of tools are particularly important (or, at least, attract considerable attention from those concerned with developing data mining tools). They include tools for unsupervised classification (clustering), supervised classification, more general predictive models (regression), modelling time series to detect trend and other structures, and graphical models.

Even though the objective and basic nature of tools for modelling will be familiar, the algorithms may differ from those in standard use in statistical applications (recall the comment about sequential estimation methods in Section 1), and often the emphasis is different. For example, recursive partitioning methods are widely used in data mining applications with emphasis placed on their interpretability. This perhaps has more in common with machine learning work than statistics. Often, in fact, the data mining work goes further in this direction and the aim is not so much to extract entire tree structures as to characterize "local rules" relating variables (e.g., "If A is present and B is absent, then C has a high probability of being present").

In some contexts, rule extraction, (or *association analysis*, as it is sometimes rather unfortunately called) is a key aim. A classic example is *market basket analysis*, so called because of its origins in mining supermarket purchasing data: interest lies in the percent of customers who purchase certain goods, given that they purchase others. Some subtleties were not always recognised by early workers in the field (or, dare we say it, by some more recent workers), often working in ignorance of the statistical content of the analysis. In particular, it will be obvious to statisticians that a high conditional probability $P(X|Y = 0)$ may not be very interesting in itself. It may only become interesting if, also, $P(X|Y = 1)$ is low, so that a contrasting behavior between groups is evident. Likewise, the familiar caution that correlation does not imply causation has not always been remembered in the enthusiasm inspired by the discovery of apparent relationships between purchases of different goods. The fact that most people who buy A also buy B does not mean that inducing other people to buy A will lead to them also buying B.

The term *data visualization* is often used to describe graphical methods in data mining contexts, where the potentially huge numbers of data points can present problems for standard statistical methods. As with statistics, dynamic and interactive methods obviously have considerable novel potential. This is an active area of research, though thus far highly innovatory tools with wide impact seem few and far between. Virtual reality methods for exploring large databases represent one exciting area of future work. Given the difficulty of adequately demonstrating dynamic, interactive and virtual reality methods in the medium of a journal paper, we will not dwell on them here. Examples of data visualization methods are described in

Cox, Eick, Wills and Brachman (1997), Derthick, Kolojejchick and Roth (1997) and Mihalisin and Timlin (1997).

Pattern detection methods will be less familiar to statisticians than will model building exercises. They represent a relatively new area of work. Much of it is rather ad hoc, and deep ideas about the best strategies to pursue in pattern detection appear not to have emerged yet. (Of course, they may never emerge.) Sections 2.1 and 2.2 give some relatively straightforward, but we hope interesting, examples of model building and pattern detection in data mining from our own work.

2.1 Models

Figure 1, from Hand (1998b), shows a histogram of the number of weeks of a particular year that credit card holders used their cards in supermarkets. This distribution has various features, which are evident from the histogram and which we might try to model. Most strikingly, there is the rapidly decaying left-hand mode. This mode is probably to be expected: it means that many people never use their cards in a supermarket; a smaller number, though still relatively large, use them in just one week and so on. However, closer examination shows that the distribution does not decay to zero, but perhaps to some constant level: a similar number of customers use their cards 20 weeks as do 30 or 40 weeks of the year. Furthermore, at the right-hand end of the distribution, there appears to be a smaller, flatter mode. Having detected this second mode, it is not difficult to concoct a retrospective explanation. It is likely that there are people who use their cards in a supermarket every week, except, perhaps, when they are on holiday or unable to for some other reason. One could build a model (perhaps a mixture model involving a Poisson left-hand mode, a reversed Poisson right-hand mode, and maybe other components) to summarize the shape of the distribution in terms of a few convenient parameters. Whether the features of this model (the decaying left-hand mode, the constant middle part, the small right-hand mode) are interesting or valuable is a matter for the supermarket operators to determine. We could conduct statistical significance tests on aspects of the model, though the numbers involved in this example are such that all the aspects above will be highly significant. Moreover, as noted above, one has to be careful in interpreting such tests; the data may be all credit card supermarket transactions in the year in question, not a sample from them. Statistical inferences will then presumably refer to other data sets which could have been drawn, with the implication that one is really concerned with future possible transactions.

This example illustrates the basics of the modelling approach to data mining. It is very similar to standard statistical modelling. Perhaps one difference is that one might be more sensitive to the question of how large ("substantively significant") are the features one wants. If the data set involved 100 million points, then minute features could be detected as highly statistically significant. It is likely, however, that many of these would be so small as to be of no conceivable value. Of course, such issues depend critically on the context. One

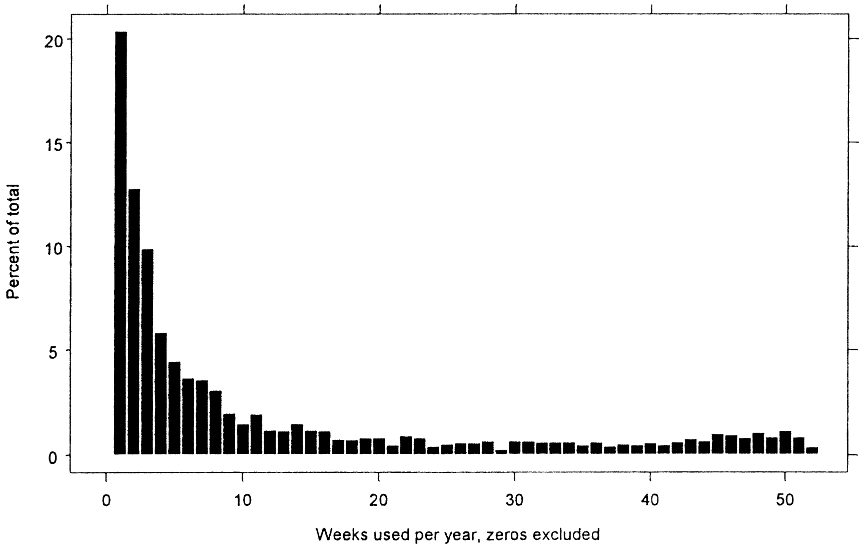

FIG. 1. *Histogram of number of times credit card owners used their cards in a supermarket in one year.*

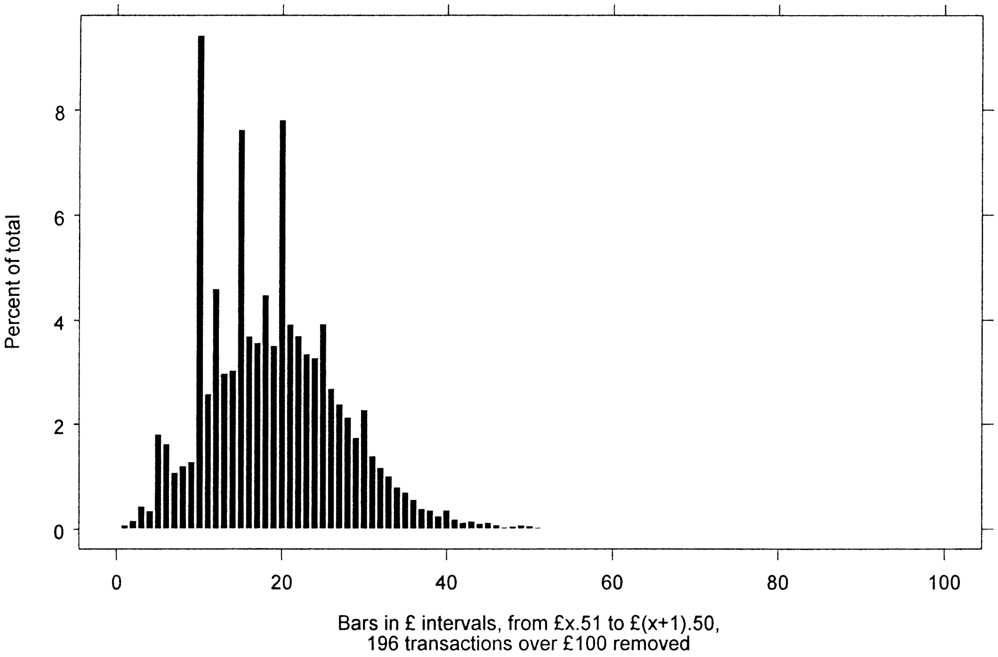

FIG. 2. *Histogram of sizes of petrol station transactions.*

can avoid detecting small features as statistically significant by reducing the size of the data set by sampling. Then standard statistical inferential issues come into play.

For a second modelling example, consider the distribution shown in Figure 2. This again involves credit card data, but now is a histogram of the amount spent in each of 40,068 credit card transactions in petrol stations over a given period. The cells in this histogram range from £x.51 to £$(x+1)$.50, and 196 transactions of value over £100 have been excluded so that the distribution of the remainder is clearer. It is apparent that the distribution has some striking and, at least to the authors, unexpected features. There are marked peaks at multiples of £5 and £10, and also at the values £12 and £18. Again the numbers of transactions here, as well as the regularity of the patterns, make it clear that these structures are not simply chance events: something real is going on. (We note here that the peaks could legitimately be regarded as patterns. In this example, however, we are concerned with modelling them, rather than detecting them. In the next section we give an example where the emphasis is on detecting and explaining patterns forming similar peaks.)

Blunt and Hand (1999) have examined this data set in detail. Closer examination reveals rounding to all integral numbers of pounds, though much less marked than is evident in Figure 2. Blunt and Hand build a mixture model for these data, in which petrol purchasers are characterized as being one of two types: those who seek to make purchases of rounded values and those who do not. Based on this, they identify features which distinguish between the two types. For commercial reasons, they stop short of showing how this information may be made use of by the credit card company.

Predictive models are important in data mining, perhaps especially in commercial applications, where one is often concerned with the possibility of manipulating practices so as to increase sales. In one of our studies we knew that customers who exhibited a certain type of transaction pattern were likely to respond positively to a marketing initiative, while those who did not exhibit the pattern were unlikely to respond positively. We knew that, in the customer base in general, 12.5% of customers followed the transaction pattern. However, we were able to identify a subgroup of 10% of the customer base that had a 43% chance of exhibiting the transaction pattern. It was then easy to set up an equation including per capita return on the marketing initiative and per capita marketing cost, to demonstrate that greater profit would be obtained by restricting the marketing initiative to the identified subgroup. Here we knew what transaction pattern was to be used as the predictor variable, so we only had to concern ourselves with the modelling aspects. In general, we will also be searching for patterns that give us this sort of opportunity. This is described in the next section.

These examples show that modelling in data mining is closely related to modelling in statistics. There are differences, however, primarily arising from the sizes of the data sets frequently encountered in data mining. These issues are discussed in Section 3. Perhaps it is not necessary to say that the border between the two disciplines, as far as modelling work goes, is a fuzzy and shifting boundary, and one that probably depends as much on the individual investigator as the subject matter and objective.

2.2 Patterns

A "pattern" is an unusual structure or relationship in the data set. The structure may be shared by relatively few cases, but nevertheless enough to be worthy of attention. We should note that the term "pattern" is used in a different way from its use in the phrase "pattern recognition." There it refers to the identification of a particular shape (in an image) or classification of a vector of observations (in statistical pattern recognition).

There are different kinds of patterns. They may be cases that demonstrate apparent departure from models, such as outliers. They may be shapes in time series or patterns in event sequences which occasionally recur (these are often called *episodes*: see Mannila, Toivonen and Verkamo, 1997). Our first example illustrates a case in which they are defined relative to a notion of uniformity of the background data.

Adams and Hand (1999) describe a tool for detecting local groups of points which might be regarded as anomalously similar. If they are not merely chance aggregations of points, such structures would reflect local peaks of the underlying distribution. Multivariate nonparametric density estimation techniques (e.g., Scott, 1992) can be used to smooth a data set to detect such peaks, but finding their positions after the smooth is difficult in more than two or three dimensions. To overcome this difficulty, Adams and Hand (1999) evaluate the smooth at each of the data points themselves. The value of the estimated density at each data point is then compared with the estimated value for the point's M neighbors. If it has a larger estimated value than any of its M neighbors, then it is taken as a local peak, and the position flagged as a pattern of potential interest. By varying M we vary the meaning of "local."

To illustrate, the left-hand panel of Figure 3 shows a scatterplot of two standardized variables measured on 5520 human chromosomes. The right-hand panel shows the position of the local maxima of estimated probability density, using $M = 10$. There are clear local maxima around the "center" of the plot, which are to be expected. However, there are other, less predictable and far from obvious peaks in the tails of the distributions. The local relative overabundance of points at these coordinates may merely be chance fluctuations, but we are flagging them as worth checking. Of course, the value of such an exercise is more apparent when more than two variables are involved.

Pattern detection data mining tools are used for identifying fraud, fault detection, instrument breakdown, distinguishing between genuine and spurious alarms, and, of course, also for finding errors in data. Hand (1998b) notes that credit card companies use such methods to trigger an alert when the pattern of card usage deviates from the customer's normal pattern.

We pointed out in Section 2.1 that sampling can be used in modelling, where the aim is generally to characterize the most important (and hence, large scale) features of a data set. However, sampling cannot be used in data mining for patterns. By definition, sampling thins out the available data for detecting the structure, and, since a pattern may be

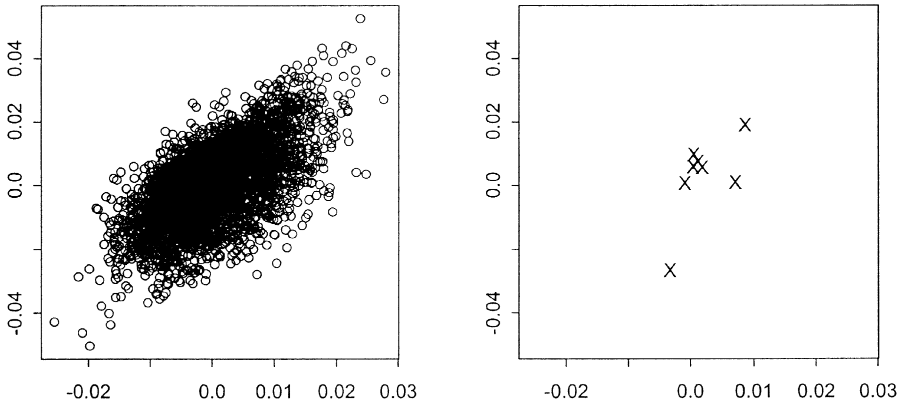

FIG. 3. *Scatterplot (left-hand panel) of chromosome data showing locations (right-hand panel) of locally dense regions.*

based on only a small number of cases in the first place, reducing this number may remove the pattern altogether. An increasing trend in a customer's number of credit card transactions may be adequately modelled by taking only every tenth transaction, but the fraud patterns mentioned above could well be missed by thinning out the data in this way.

There are two major problems with pattern detection methods. One arises from data quality and is discussed in Section 4. The other, familiar to statisticians, is the fact that many spurious patterns, generated merely by chance fluctuation, will be "discovered." This is particularly the case if one is examining many candidates, seeking patterns defined by small numbers of cases. We discuss this further in Section 3. Whereas much statistical work is concerned with characterizing how likely it is that apparent structure will arise in a data set given that there is no such structure in the underlying process, in data mining the emphasis is simply on *locating* the structure. The responsibility for deciding whether the structure has meaning in terms of the underlying process ("is real") is shifted to a domain expert. For example, perhaps the 100 "largest" structures in the database will be referred to the expert, who can decide whether they are interesting, valuable, or likely merely to be chance features of the random aspect of the data. This is one reason why "data mining" should not be seen as a "once off" activity, but rather as a process, in which the discovery of apparent structure and the interpretation of that structure bounce back and forth.

Although the aim of the exercise is to find previously unsuspected patterns, we believe that patterns which can be explained (that is, for which a convincing post hoc rational explanation can be created) are much more likely to be "real" than those for which such an explanation cannot be created. Thus, we suggested an explanation for the right-hand mode in Figure 1 that, at least to us, sounds reasonable. It would be more difficult to find an equally convincing explanation for the mode at 11 weeks in that figure. Although a pattern detection data mining tool [such as that described by Adams and Hand (1999), outlined above] could well pick up this mode also, we would be suspicious of it on the grounds that we could not explain it.

Figure 4 shows a histogram of the values of credit card purchases in department stores. There are clear peaks at some values; notably at values ending in 0 and, to a rather lesser extent, at values ending in 5. A pattern detection algorithm, such as that described above, might pick these up (though, in fact, since this example is one-dimensional for illustrative purposes, this is not really necessary). Since department store purchases seldom provide the opportunity for choosing rounded values, the explanation provided for the petrol purchases in Section 2.1 will not suffice. Neither will digit preference (see Section 4) since the values are objective

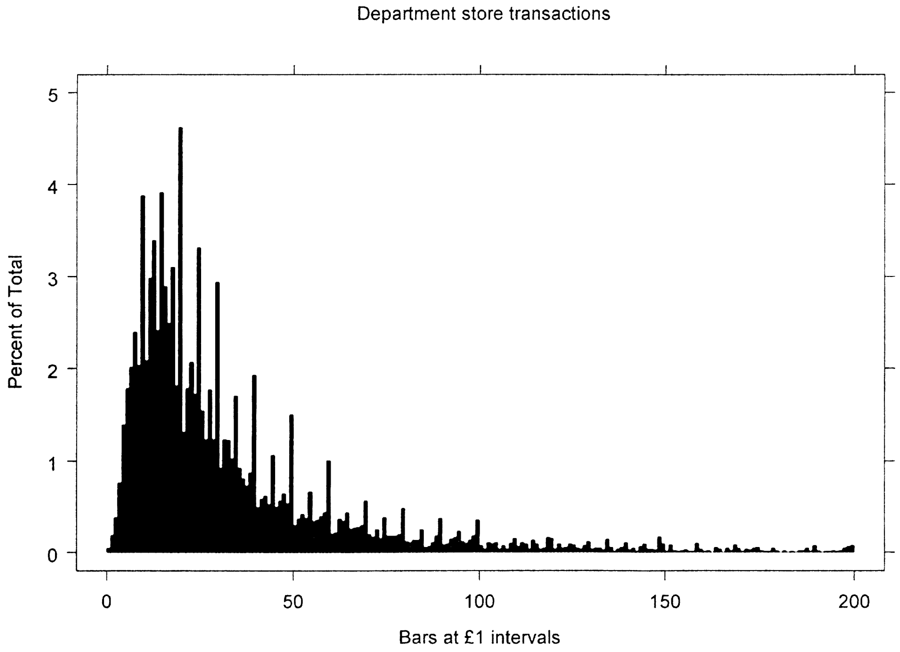

FIG. 4. *Values of credit card purchases in department stores:* £1 *cell widths.*

and recorded electronically. Some other explanation is required and a little thought readily provides one.

Items in department stores are often priced to 1p less than a multiple of 5 or 10 pounds. Thus we have prices of £4.99 and £19.99. If plotted on a histogram with cell widths of £1, these will be rounded to £5 and £20. Further support for this explanation comes from Figure 5, which shows histograms, with cell widths 1p, around some of the peaks from Figure 4. If one, or sometimes more, purchases with such prices were made, one would expect to find precisely these local distributions, tailing off to the left.

The structure in Figure 4 can be explained, and we are confident that (a) the structure is real and (b) our explanation is correct. Whether the information is valuable or not is a different question. Whether the structure is obvious or not is also a different question; perhaps most of the real structures discovered in data mining exercises are obvious *in retrospect*.

Pattern detection is often used in an interactive way. Thus *collaborative filtering* describes the process of extracting conditional probabilities from purchasing data used by some retail organizations. (For example, the *Amazon* internet bookstore uses this strategy.) The purchases of customers who bought item X are examined to see what other items, Y, they tended to buy. Then new customers who buy X are alerted to the fact that they may also find Y interesting. Clearly, rather more than a high conditional probability $\mathrm{P}(Y|X)$ is needed—some items are bought by everyone.

The key feature is a high $\mathrm{P}(Y|X)$ and a low $\mathrm{P}(Y|{\sim}X)$, along with, of course, a value of $\mathrm{P}(X)$ which is not too small. The exercise here is a fairly elementary statistical one, although the practice does have some interesting features. The data sets are large, of course, and the application (detect a purchase of item X and send a message about Y) occurs in real time, although the processing to find the conditional and marginal probabilities need not. On the other hand, these probabilities need regular updating, and this updating must also adapt to the changing inventory of the supplier.

3. PROBLEMS OF SIZE

It is the size of the data sets encountered in many data mining applications which provides the potential for discovering and taking advantage of novel structures. The sheer number of records may serve to conceal features of interest, but these numbers mean that relatively small effects, not easily identifiable or observable with smaller numbers, can be productively sought. But size is a two-edged sword. As well as providing the opportunity, it also brings with it problems.

One obvious one is that large samples mean that things which are substantively insignificant may

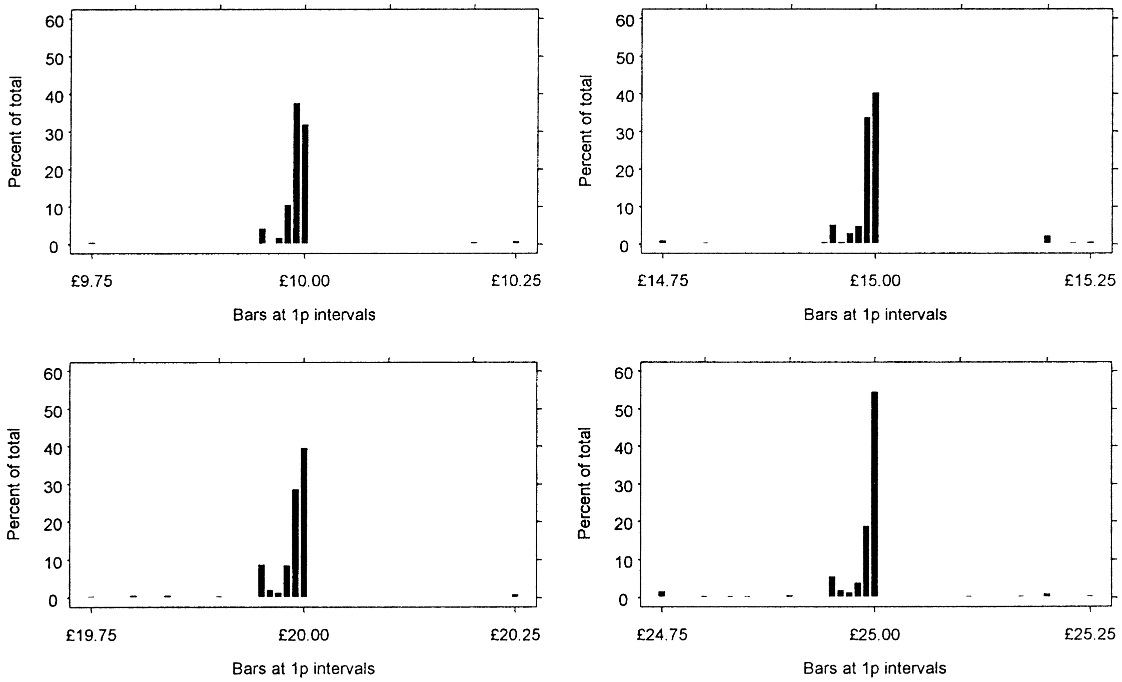

FIG. 5. *Histograms of credit card purchases in department stores:* 1p *cell widths.*

be highly statistically significant. This is an issue which is very much context dependent. For example, 100,000 records with a common property may mean very high statistical significance. And 100,000 records may provide the potential for large extra profit, if advantage can be taken of whatever common property they have. (In one project with which we have been involved, reducing the bad rate on a bank loan portfolio of 8 million customers by just 0.25%, in which each "bad" means a loss of £1000, would lead to an overall saving of £20 million.) On the other hand, if the 100,000 records are from a 1 billion item database, then the relative size of the extra profit, may not be worthwhile. It depends on the context.

In searching for interesting patterns, one could test each candidate pattern for statistical significance. However, this inevitably means a large number of tests. Carrying out large numbers of tests does not really protect us against detection of spurious relationships, either because they will also falsely reject a large number of null hypotheses (5% of a large number is still a large number, for example) or because the overall ("experimentwise") control adopted requires the underlying effect to be very large to have a nonnegligible chance of being detected.

Given that formal probability models are of limited value, a *scoring* strategy is often adopted. This abandons probabilistic interpretations and simply scores each model or pattern according to its "unusualness," "unexpectedness," or "interestingness." The quotation marks are intended to signify that there is some considerable flexibility in defining the measures used for these concepts. (Note the similarity to projection pursuit, though there the aim is really modelling rather than pattern detection.)

Other issues arising from the size of modern data sets relate to how to analyse them. For example, the scatterplot is a basic statistical tool that is very useful both for probing data and for communicating findings. However, to illustrate what can happen when very large data sets are involved, Figure 6(a) shows a scatterplot of a data set of time in employment against the day of application (called "index" in the plots) for around 96,000 bank customers. The scatterplot does show some structure. For example, there are clear dense horizontal bands. However, it is likely that these bands are an artifact of the coarseness with which the data are recorded, and do not reflect any underlying reality of interest. In fact, this scatterplot is concealing, rather than revealing information, as the contour plot of the same data given in Figure 6(b) shows. There are clear modes for both time in employment and index number, though there appears to be no dependence between the values of these two variables.

This example illustrates a key issue for data mining algorithms: they must be *scalable*. That is, they must increase in a reasonable way as the sample size (or perhaps as the number of variables) increases. This scalability must relate to the ease with which the methods reveal aspects of structure, and also to computational resources. Algorithms in which processing time or required memory increase exponentially or quadratically with sample size rapidly become impracticable. This necessary feature of data mining algorithms may mean that methods that are optimal from a formal statistical perspective cannot be used.

4. DATA QUALITY

We believe that data quality is one of the key issues in data mining. The simple reason is that, as any experienced data analyst will know, it is extremely unusual to find data sets that have no errors or distortions. Presented with a data set that appears to be free of errors, a statistician's suspicions may be immediately aroused. One might ask what happened to the incomplete records, whether the recording instruments really never failed or drifted over time, if all the cases were followed to the end of the study period, and so on.

This being the case, and given that data mining is concerned with finding structure in data, we might reasonably expect such exercises often to identify structures arising solely because of inadequacies or peculiarities in the data or the data collection process. Our experience shows this expectation to be entirely justified. Figure 7 shows a histogram of diastolic blood pressure for a sample of 10,772 men. A crude analysis would reveal that there is a strong correlation between the value of the blood pressure and whether or not it takes odd-numbered values. Indeed, the histogram shows that there are several other peculiarities with this data set. For example, there are marked peaks at values ending in 0, and the only odd values which occur in the lower part of the distribution are those ending in 5. We attribute the peaks to digit preference (contrast this with the explanations for the peaks in the petrol purchase data and the department store data above). Moreover, deeper investigation revealed that the instrument only recorded even values and, when measurement yielded an exceptionally high value of blood pressure, the measurement was repeated and the average of the two (even) measurements recorded. This can yield odd numbers.

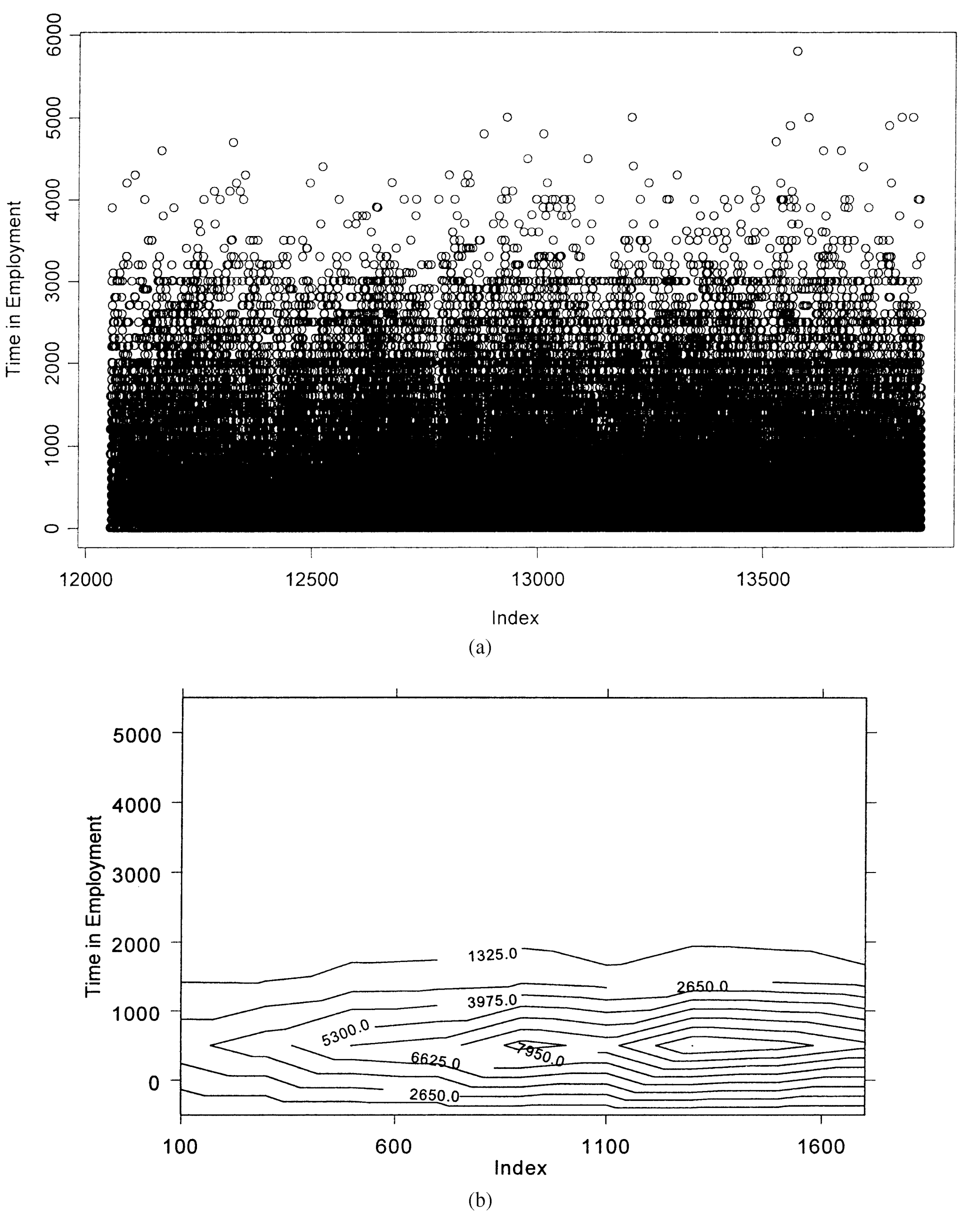

FIG. 6. (a) *Scatterplot of* 100,000 *bank customers, showing their time in employment plotted against their index number.* (b) *A contour plot corresponding to Figure* 6(a).

We believe the values ending in 5 in the lower part of the distribution are again due to digit preference. Of course, the necessity for seeking such explanations only arises because one has identified unexpected features about the data in the first place.

We believe it is not farfetched to suggest that most of the "interesting and unexpected" patterns discovered in a data set during the course of a data mining exercise will be attributable to "inadequacies" of the data.

Data distortion and missing values can, of course, be a serious problem even in relatively small data sets. Figure 8 is produced from a data set of 3884 applicants for loans, with the application form yielding scores for 25 variables. The histogram shows how many applicants had no missing values, one missing value and so on. Only 66 applicants provided complete information, and one applicant had 16 of the 25 values missing (history does not record whether this applicant was granted a loan). Only five of the variables had no missing values, and two had over 2000 missing values.

Of course, things are complicated by the fact that, often, whether a value will be recorded for a variable is contingent on the values taken by other variables. If half the subjects are female, questions about testicular cancer remain unanswered. This sort of sit-

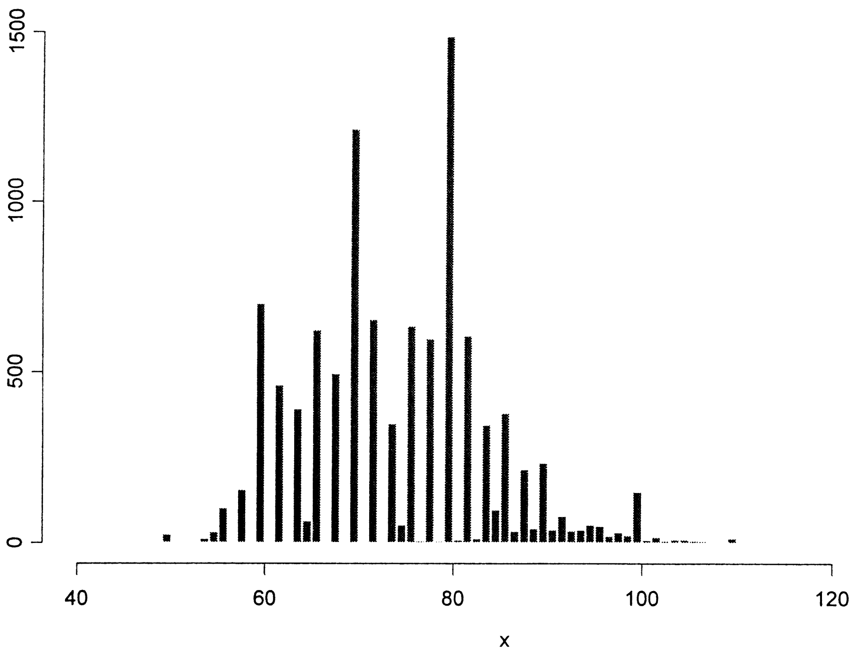

FIG. 7. *Diastolic blood pressure for* 10,772 *men.*

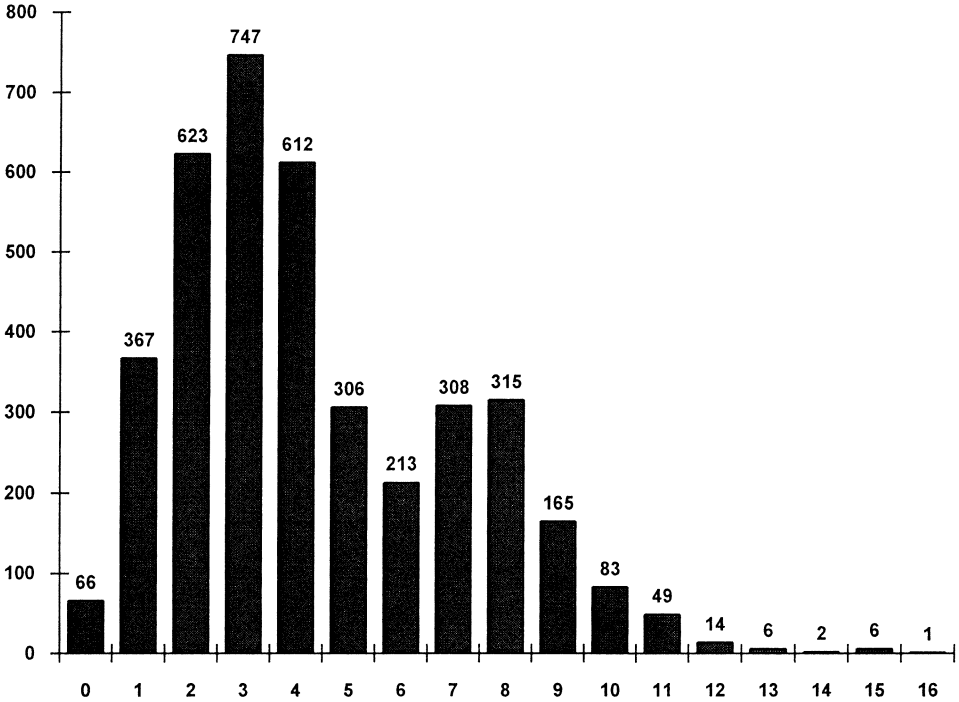

FIG. 8. *Histogram of number of missing values for each of* 3884 *applicants for personal loans.*

uation is illustrated in Figure 9. The data are from a study to develop a screening instrument for osteoporosis and relate to 1012 elderly women (Cooper, et al., 1991). The vertical axis, V1, shows the patient number and the horizontal axis shows the variable number (from 1 to 45). A point is printed whenever there is a missing value. The resolution of the figure is such that one cannot see individual cases, but it is apparent from the figure that there is structure to the missing values. Variables 4 and 5 are usually either both missing or both not missing. A few cases have most variables missing. Some variables have many missing values. And so on. Many such features are precisely the sorts of things one would hope a data mining exercise would detect, but they will often be of little interest; it is typically statements about the values that *are* recorded that one is seeking to make. (On the other hand, missing values can sometimes be useful; for example, in supervised classification problems, sometimes the fact that a value is missing can be predictive.)

Missing values are perhaps the most common kind of data distortion, but there is an infinite number of ways that things can go wrong. Figure 10 shows mean weight changes over a four-month period for the control group from a classic study of 10,000 children carried out in the 1930s to explore the effect of free school milk (Leighton and McKinlay, 1930). The consecutive pairs of points show the mean weight before the trial and the mean weight at the end of the four-month trial. The different pairs of points relate to different age groups. A priori, one might expect that the curve would be smooth. The manifest irregularities are precisely the sort of thing one would hope a data mining exercise would detect, such structure is unexpected, and certainly interesting. But in this case it is of little value. It seems likely that it can be attributed to the fact that the first measurement of each pair took place in February and the second in June when the weather was warmer and the children would be wearing lighter clothing. (This despite the precautions the experimenters took: “All of the children were weighed without their boots or shoes and wearing only their ordinary indoor clothing. The boys were made to turn out the miscellaneous collection of articles which is normally found in their pockets, and overcoats, mufflers, etc., were also discarded. Where a child was found to be wearing three or four jerseys—a not uncommon experience—all in excess of one were removed” (Leighton and McKinlay, 1930, page 8).

Hand (1998b) distinguishes between two kinds of data inadequacy. In one kind, individual records are distorted, while, in the other kind, the overall sample is distorted. If individual records are distorted, one might try to develop methods that allow for this or correct the distortion. Thus, for example, the EM algorithm will permit one to find maximum likelihood solutions with incomplete records and impu-

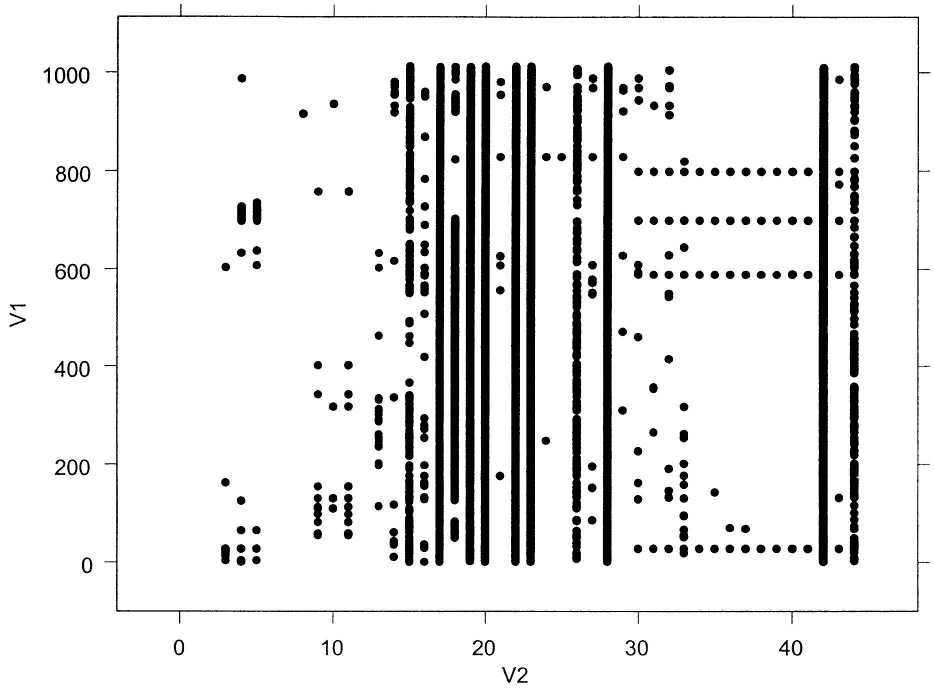

FIG. 9. *A missing value plot of* 1012 *elderly women.*

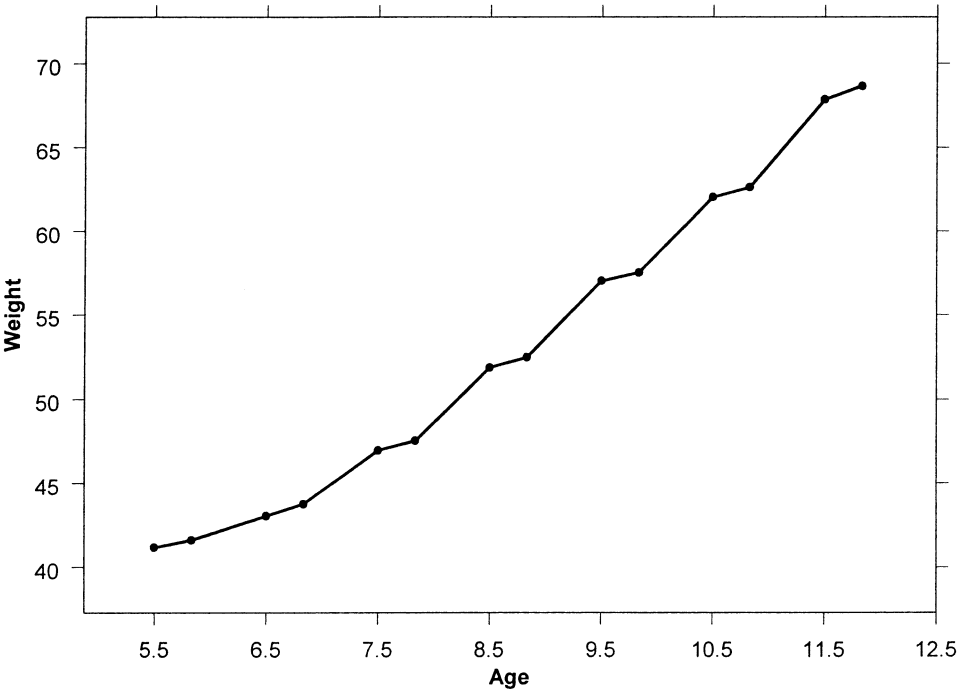

FIG. 10. *Weight changes for* 10,000 *children in a trial to study the effect of milk on growth carried out in the* 1930*s.*

tation methods seek to fill in the missing values. Of course, such methods are based on assumptions and models about why the data are missing. The lack of transparency which is an inevitable concomitant of very large data sets means that one may be tempted to adopt automatic "data correction" procedures. While these may be useful, there is also the very real risk that they will smooth out or remove the very structures one is hoping to detect. An anomalous tight grouping of a handful of customers could be regarded as due to data recording errors. Clearly the solutions are not straightforward.

The second kind of data inadequacy occurs at a higher level and is due to distortion in the set of records which is selected for inclusion in the database. Selection bias is an example of this sort of phenomenon, in which whether or not a record is included in the database depends on the values the variables take. This can cause major difficulties. Patients for whom a treatment fails to work may be more likely to drop out of a study, thus leaving those for whom it has worked and hence giving a distorted impression of the treatment's efficacy. Copas and Li (1997) give an example involving kidney dialysis, in which log(hospitalization rate) for a new method of treatment appears to decline with time, but for which it is possible that the effect is due to the nonrandom allocation to the new and standard treatment.

In many data mining problems, in which data collection may be a far from straightforward process, it is likely that the samples in the database will be convenience or opportunity samples: those records which could be easily collected. The implications for inference to the overall population may be unpredictable. Furthermore, populations can change over time. Figure 11 shows weekly averages of four variables describing applicants for unsecured personal loans over a four-year period (time-scale disguised for commercial reasons). Figure 11(a) (a binary indicator of whether or not customers were aged between 30 and 35) shows very little change over time. Figure 11(b) (a binary indicator of whether or not the customer had a check guarantee card) shows a marked downward trend. Figure 11(c) (a binary indicator of whether or not the loan was for debt consolidation) shows a gradual upward trend, with a superimposed seasonal component. Finally, Figure 11(d) (a binary indicator of repayment method) shows a clear change halfway through the observation period. This is thought to be due to a policy change by the bank, requiring customers to pay by a particular method, and is perhaps therefore an example of an uninteresting structure. (On the other hand, one might enquire why, if customers are *required* to pay by a particular method, there is any irregularity at all in the right-hand half of the figure. Perhaps this is another issue of data quality.) In such cases, inferences made on data collected at one time may have

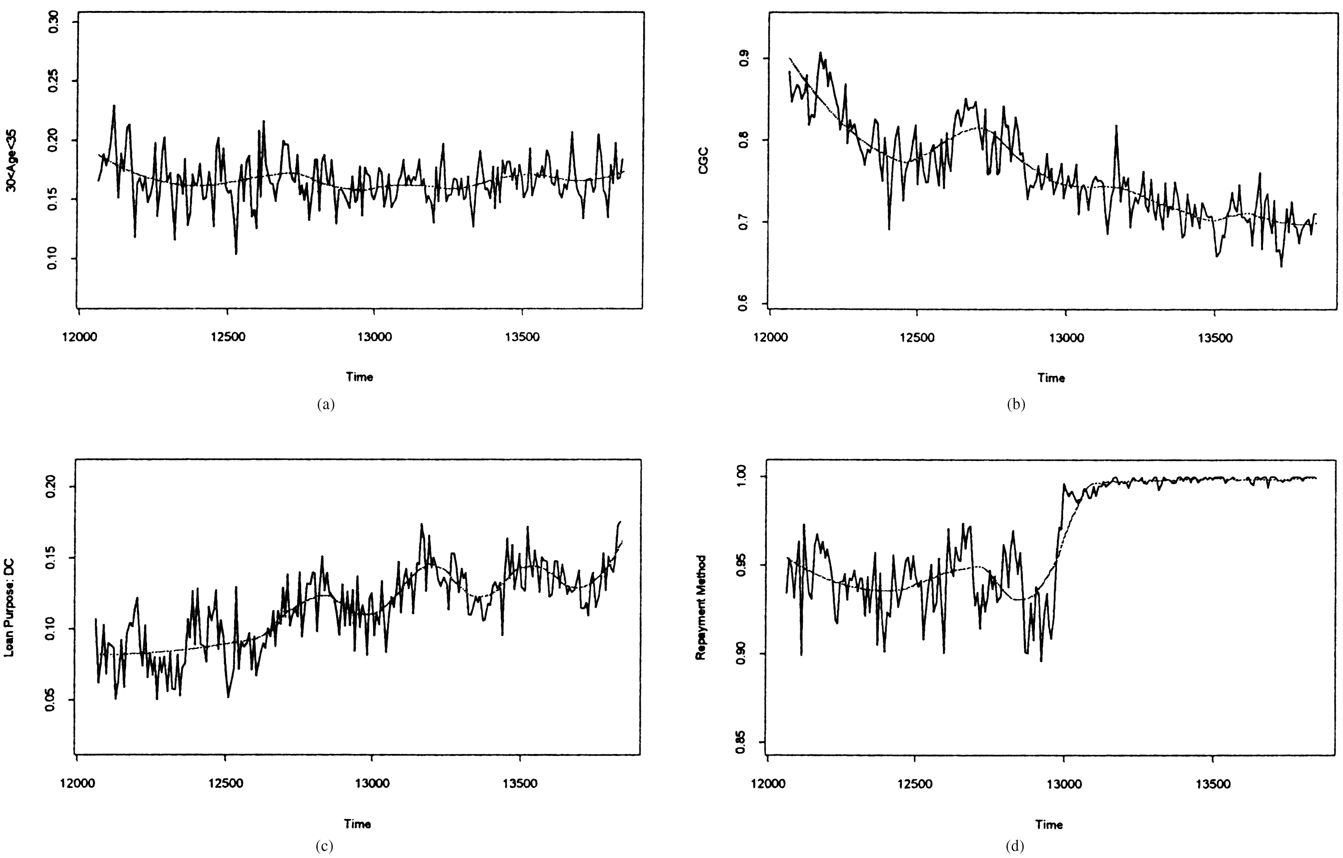

FIG. 11. *Weekly averages of four binary variables.* (a) *Customers aged between* 30 *and* 35 *years,* (b) *Check guarantee card,* (c) *Purpose of loan—debt consolidation,* (d) *Repayment method.*

limited applicability later on. Of course, in such circumstances one might contemplate building a dynamic model.

As a final example, consider Figure 12. This arose during a study of how well different definitions of a “bad” risk could be predicted in a consumer banking operation. The vertical axis shows the Gini coefficient, a measure of how well a prediction rule performs. The horizontal axis shows different values of a threshold, varying between 1 and 27, so that different definitions of bad risk result. There is a striking pattern in the plot, as in several of the examples above, precisely the sort of pattern one might hope a data mining exercise would identify. Unfortunately, once again, this pattern is an artifact of problems with the data. In this case, the bank in question had previously worked with a definition in which values of the horizontal variable above 5 were taken to define a bad risk. This variable had been inadvertently included as a potential predictor in the analysis, and the looping pattern in the plot is a direct consequence of this.

For explanatory purposes, each of the examples above has focused on single problems. But life is seldom so simple. It is far more likely that the sample will be distorted, *and* there will be missing values, *and* those values which are recorded will sometimes be misrecorded and so on. In one data set on unsecured personal loans, we found tiny amounts unpaid (e.g., 1p or 2p) leading to a customer being classified as a “bad debt,” negative values in the amount owed, 12 month loans still active after 24 months (technically not possible under the bank’s rules), outstanding balances dropping to zero and then becoming positive again, balances which were always zero and number of months in arrears increasing by more than a single integer in one month. Once identified, some of these problems can be explained (and, perhaps, corrected). Some, however, defied ready explanation.

5. CONCLUSION

The large size of modern data sets means that it was legitimate to coin a new phrase to denote the activity of finding structure and pattern in data. Modern statistics has grown from a base of analysis of relatively small data sets. Moreover, modern statistics has placed emphasis on only parts of the entire range of data analytic problems (see, for

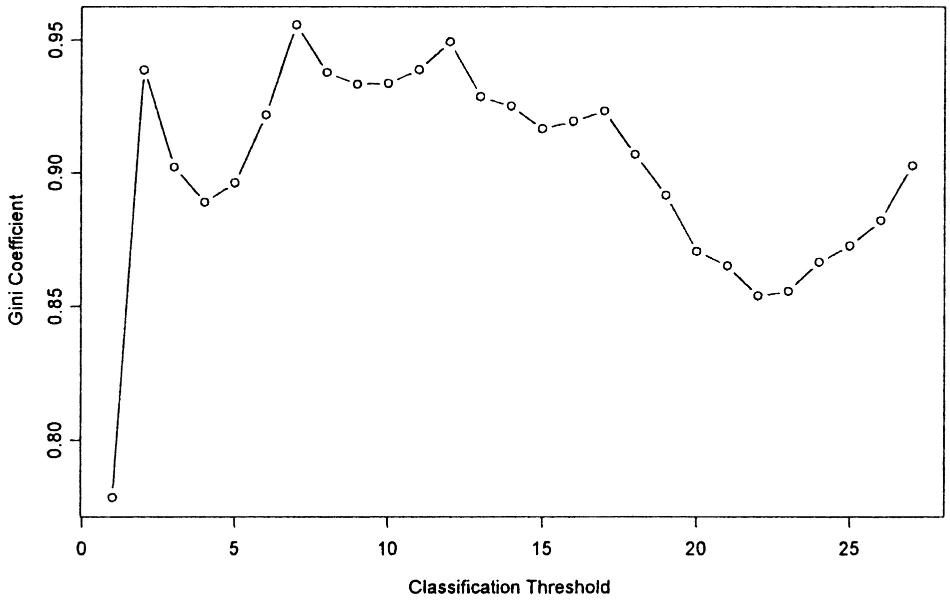

FIG. 12. *Structure induced in classification rule performance measure due to inclusion of an inappropriate variable.*

example, Chambers, 1993 and Bartholomew, 1995) and other disciplines, especially machine learning and database technology, have begun to make important contributions.

Because data mining is such a young field, there remain fundamental open questions, requiring novel ideas for their solution. Among the most important of these questions are those relating to issues of chance in the context of large numbers of events and to issues of data quality.

Also because it is a new field, data mining has developed its own terminology (for example, dicing, drilling down, rolling up, and support and confidence). In many cases, there already existed perfectly good terms for the same exercises in disciplines such as database theory or statistics (conditioning, marginalizing, joint probability, conditional probability). In other cases, terms have been invented according to the domain in which they were developed. “Market basket analysis,” mentioned above, is one such, as is *bread dominoes*, describing how records can chain together (derived from the fact that shoppers unable to buy the kind of bread they want tend to purchase some similar alternative).

One important difference between data mining and statistics is the emphasis of the former on algorithms. In this regard, data mining has more in common with machine learning than statistics. We would go so far as to suggest that the data mining literature is full of descriptions of algorithms with little underlying theory relating them. It is easy to develop new algorithms, but without underlying theory it is difficult to see what sort of progress is being made. It is possible that, as the field matures, so deeper theoretical understanding will lead to proper critical assessment of the algorithms and their (comparative) properties. However, it is also possible that the ease with which algorithms can be developed using powerful computers will work against that.

Data mining has promise, but there are many difficulties associated with it. It is not to be entered into lightly or in ignorance of the obstacles. Perhaps there are similarities to meta-analysis, in that it is easy to carry out a poor analysis, and very hard to carry out a good one. One difference is that the quality of the data mining exercise will soon be revealed, in terms of whether the structures which have been unearthed are interesting, valuable, surprising or previously unknown. Also like meta-analysis, the phrase “data mining” is rich in implication. It almost spells excitement and opportunity. However, given the difficulties we have outlined above, one should be wary of getting carried away by this. It remains to see precisely who will benefit from data mining activities, beyond companies marketing data mining tools.

General overviews of data mining, including discussions of the relationship between data mining and statistics, are given in Elder and Pregibon (1996), Fayyad, Piatetsky-Shapiro and Smyth (1996), Glymour, Madigan, Pregibon and Smyth (1997), Klösgen (1998), Hand (1998a, b), Hand (1999) and Hand, Mannila and Smyth (2000). The range of current research is showcased in the journal *Data Mining and Knowledge Discovery* and the proceedings of the *International Conference on Knowledge*

Discovery and Data Mining series (e.g., Heckerman, Mannila, Pregibon and Uthurusamy 1997 and Agrawal, Stolorz and Piatetsky-Shapiro, 1998).

ACKNOWLEDGMENTS

We are grateful to Littlewoods Plc, Abbey National Plc, Barclaycard and Fergus Daly for permission to use their data to illustrate some of the points made in this paper.

REFERENCES

Adams, N. M. and Hand, D. J. (1999). Mining for unusual patterns in data. Working paper, Dept. Mathematics, Imperial College, London.

Agrawal, R., Stolorz, P. and Piatetsky-Shapiro, G. (eds.) (1998). *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA.

Babcock, C. (1994). Parallel processing mines retail data. *Computer World*. Sept. 6.

Bartholomew, D. J. (1995). What is statistics? *J. Roy. Statist. Soc. Ser. A* **158** 1–20.

Blunt, G. and Hand, D. J. (1999). Credit card petrol purchases: an example of data mining in practice. Working paper, Dept. Mathematics, Imperial College, London.

Chambers, J. M. (1993). Greater or lesser statistics: a choice for future research. *Statist. Comput.* **3** 182–184.

Cooper, C., Shah, S., Hand, D. J., Compston, J., Davie, M. and Woolf, A. (1991). Screening for vertebral osteoporosis using individual risk factors. *Osteoporosis International* **2** 48–53.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples. *J. Roy. Statist. Soc. Ser. B* **59** 55–95.

Cortes, C. and Pregibon, D. (1997). Mega-monitoring. Paper presented at the Univ. Washington/Microsoft Summer Research Institute on Data Mining, July 6–11.

Cox, K. C., Eick, S. G., Wills, G. J. and Brachman R. J. (1997). Visual data mining: recognizing telephone calling fraud. *Data Mining and Knowledge Discovery* **1** 225–231.

Derthick, M., Kolojejchick, J. and Roth, S. F. (1997). An interactive visualisation environment for data exploration. In *Proceeding of the Third International Conference on Knowledge Discovery and Data Mining* (D. Heckerman, H. Mannila, D. Pregibon and R. Uthurusamy, eds.) 2–9. AAAI Press, Menlo Park.

Elder, J. IV and Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds.) 83–113. AAAI Press, Menlo Park, CA.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds.) 1–34. AAAI Press, Menlo Park, CA.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery* **1** 11–28.

Hand, D. J. (1998a). Data mining: statistics and more? *Amer. Statist.* **52** 112–118.

Hand, D. J. (1998b). Data mining—reaching beyond statistics. *Res. Official Statist.* **2** 5–17.

Hand, D. J. (1999). Statistics and data mining: intersecting disciplines. *SIGKDD Exploration* **1** 16–19.

Hand, D. J., Mannila, H. and Smyth, P. (2000). *Principles of Data Mining*. MIT Press.

Harrison, D. (1993). Backing up. *Neural Computation* 98–104.

Heckerman, D., Mannila, H., Pregibon, D. and Uthurusamy, R. (eds.) (1997). *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA.

Klösgen, W. (1998). Analysing databases with knowledge discovery methods. *Res. Official Statist.* **1** 9–35.

Leighton, G. and McKinlay, P. L. (1930). *Milk Consumption and the Growth of School Children*. H.M. Stationery Office, Edinburgh.

Mannila, H., Toivonen, H. and Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery* **1** 259–289.

Mihalisin, T. and Timlin, J. (1997). Fast robust visual data mining. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (D. Heckerman, H. Mannila, D. Pregibon and R. Uthurusamy, eds.) 231–234. AAAI Press, Menlo Park, CA.

Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley, New York.

Comment

William Kahn

The current approach to direct marketing emphasizes data analysis and, as such, is grossly suboptimal. By approaching direct marketing in a more traditionally scientific manner, one that involves all aspects of the research process, significant new economic value can be generated. Specifically, companies will be much more successful if they use modern experimentation than by simply improving their data analysis.

1. INTRODUCTION

As an active seeker of both fun and profit, and as a statistician in the financial sector, I welcome the opportunity to comment on Hand, Blunt, Kelly and Adams. The topic is extremely timely and relevant, with tens of millions of dollars spent annually within individual financial firms and with even relatively small data mining firms being bought for hundreds of millions of dollars (*New York Times*, Oct., Nov., 1999).

However, while I appreciate the importance of data mining, in practice the profit it brings has turned out to be surprisingly limited in many key businesses. W. Edwards Deming began his 1943 statistics book, *The Statistical Adjustment of Data* with “The purpose of collecting data is to provide a basis for action.” It is in this spirit, the spirit of helping businesses make better decisions, that I wish to discuss an extension of Hand et al. beyond analysis of large data sets to the active collection of useful data.

2. THE CONTINUOUS LEARNING CYCLE

Before I jump into some of the details of research design, let me present an overview of learning which my clients have found useful. The core observation is that businesses make decisions. Some decisions, the strategic ones, can never be repeated or replicated, yet have significant impact. Examples include decisions to develop a franchise distribution model, to demutualize or to expand internationally. These decisions require leadership, commitment and insight into industry structure. Most business decisions, however, are tactical and can be—and are—repeated. Examples include how much to charge, to whom to send mailings and which product features to highlight. My main thesis is that those decisions which are repeated should be tracked and managed so as to be continuously improved.

I have broken this continuous learning cycle into eight basic operations: data archiving, financial analysis, brainstorming, experimental design, operations, data collection, data analysis and implementation. Hand et al., in focusing on just part of the data analysis issues, find themselves working with immense amounts of data which contain very little information, while trying to reflect on poorly defined business problems. A systematic approach to the entire statistical problem will greatly improve data mining’s contribution to business success. For example, one of my clients annually mails 340 million pieces of mail. One of their core business questions was to whom they should remail, that is, mail a second advertisement. Unfortunately, in all the historical data, every name from a certain source had been remailed and only those names had been remailed. There was perfect confounding between name source and remailing. No amount of data analysis on the one gigabyte data set we were given would ever allow data-driven estimation of the value of remailing. To allieviate this and related problems, we use the following eight steps to help guide our direct marketing clients.

2.1 Data Archiving

Data warehousing companies have been extremely successful over the past decade in selling multiyear, multimillion dollar corporate-wide, integrated, data warehouses. In retrospect, a large fraction of the companies investing in these projects have felt that the value generated from the investment was minimal. Terabyte databases are now common. While, as a statistician, I am happy to have data, the quest for size in these architectures is so ambitious that the usability and business value is normally left far behind. Most commonly, the warehouses support the creation of hundred-page

William Kahn is Professor, Mitchell Madison Group, West 57th Street, New York, NY 10019 (e-mail: kahnb@mmgnet.com).

weekly tallies, which hundreds of managers receive and no one actually uses to support decisions. Our experience is that wisely selected data, for example, fewer than one hundred fields per customer, turns out to have nearly all of the value and to be easily manageable. We encourage timely use of data which is immediately available with additions to the data driven by documented value and available time.

2.2 Financial Analysis

Fundamentally, business decisions are made in an attempt to improve specific business metrics. These metrics become the response variables in predictive modelling. Most companies, however, grossly suboptimize their business. It is common, for example, for a credit card company to reward the direct marketers for more credit card applications, regardless of the credit quality of the applicants. The credit evaluation group tries to minimize credit losses, but ends up weeding out extremely profitable customers who use the credit line more extensively than the original limit allowed. The servicing group is measured on minimal cost and not rewarded for effective cross-sell, and the collections group minimizes resolution time instead of total future lifetime value. In summary, each stage in the customer life cycle is optimized, based on local measures, a process which greatly suboptimizes the overall business system. It is crucial to build an integrated financial model of customer activity in order to have a proper viewpoint of long-term customer value, conditional on everything known about the customer up to that point.

2.3 Brainstorming

As in any scientific endeavor, a thorough review of all possible stimuli must be made before deciding on the key factors which will be studied. A business unit must decide which factors are under its control and could, therefore, be studied. It is thus crucial to have a view of everything a business could do. This involves explicitly knowing everything the company is trying, and has tried, both recently and long ago. Furthermore, knowledge of everything competitors are doing is obviously key. Understanding of the theoretical and academic literature, as well as of innovative and out-of-the-box approaches, will prove useful. In standard direct marketing there are dozens of potentially important factors, including the impact of color, paper quality, reading level, telemarketing, font size, mail class, delivery day, price and free gifts.

2.4 Experimental Design

Given the dozens, if not hundreds, of decisions direct marketers must make, it is not suprising that many of the decisions are based on luck and folk-knowledge. However, without the use of formal experimental design methodologies, it is impossible to learn the impact of all the decisions which must be made. Very little of the $250 billion spent every year in the United States in direct marketing is part of an experiment with complexity beyond case-control. 2^2 designs are found on occasion, but the $2^{2(5-1)}$, wherein five two-level factors and all 10 two-way interactions are studied in 16 packages, a design common in agriculture and engineering, is virtually unheard of in this field. I have run five-factor d-optimal designs in direct marketing with tremendous impact. Further, given the ability of modern in-line package production, it is now possible to customize every individual offering. I hope to run a full factorial 2ˆ15 design in the near future. While the interaction effects will likely be ignored in the analysis, given no incremental cost over the 16 offering, highly fractionated, 2ˆ(15-11) version of the design, there is no reason not to run full factorial.

2.5 Operations

In traditional scientific fields there is a constant tension, usually good-natured, between the theoreticians and the experimentalists. One component of this banter is the knowledge, by both groups, of the extreme difficulty in making laboratory equipment actually work or of making field observations reliably. In the lab, cables won't mate, and when they do, the connections are intermittent. In the field, perhaps the rainy season starts three weeks early, and the only poncho has a hole and drips rain over the camera. Similarly, in the actual execution of ten million piece (and larger) mail drops, much can, and does, go wrong. It is hard to ensure the integrity of apparently simple basic operations like color alignment, address validation and postage metering. When a theoretician asks direct mail operators to handle, say eight separate packages with correct randomization, the response is a resounding "Impossible!" After much explanation, cajoling and sometimes additional funding, the answer can be driven to "Well, OK, we'll try." However, while the eight packages are indeed actually produced, almost certainly they are mailed on successive days, and not randomly. Making experimentation work requires great attention to the capability of the operations team. It must be overseen by a principal investigator with hourly attention to detail.

2.6 Data Collection

The response to direct solicitation, whether mail, e-mail, telemarketing, in person or web, must be collected reliably and quickly. Initially, the response may just be an indicator that the targeted individual was home. Later response measures would include expressed interest in the product, order quantity, net profitability and, finally, up-sell and cross-sell producing a response equivalent to total customer-value. It is crucial to pick up each of these customer responses as soon as the data become available. I have seen large business operations ($100 million annual revenue) continue to use three-year-old response data in fields which change extremely rapidly, such as mail-order PCs. When working in a rapidly changing market, collecting timely data is crucial.

2.7 Data Analysis

Having collected timely response data from a designed experiment and having built a proper financial model for the economic value of the changed customer behavior, we are now in a position to model the change in customer value as a function of two sets of variables. The first set are the factors used in the design, that is, the factors we can actively control. The second set are the covariates which describe each customer that resides in our data archive. Note that we are not interested in which package features are best. Nor are we interested in which individuals to mail. Rather, we must model the best possible offer to mail to each individual customer. The analysis of the design factors in and of themselves is fairly simple—basic averages normally do pretty well in balanced designs. However, the introduction of the interactions of the design factors with several dozen or more covariates greatly complicates the analysis. Typically, the covariates are extremely dirty data, and robust procedures are absolutely required. Hand et al. discuss many of the issues surrounding this phase of the research. Let me add from my personal experience that with good quality design data, almost any modelling approach yields significant insights. Adding robust procedures gives significant lift, and the use of interpretable models, such as recursive trees, gives good internal saleablity.

2.8 Implementation

We are now in an enviable position. We have a formal mathematical model which gives the best possible product to offer to every individual. We also can predict the value generated by any alternative offer. Thus, we can estimate the value being generated by our optimal assignment process. This is often a startingly large number; we have routinely learned how to double the economic return in direct mailing. One interpretation of this result is that broad application of these procedures will halve the amount of junk mail received by American households, while delivering the same amount of interesting and intriguing offers. The targeting procedure is simply to mail each customer the offer which optimizes the increase in value. Of course, the null offer—do nothing—is one of the possibilities. Furthermore, because we want to be able to update our targeting models continually, we retain some small proportion of customers (say 1% to 10%) for random (as opposed to optimal) reassignment to new experiments.

3. CONCLUSION

The continuous learning cycle discussed will be very familar to readers of this journal: it is just a specific articulation of the scientific method. It is structured in a way that business managers can understand and is sufficiently detailed so as to be implemented relatively directly. While I have used direct marketing as my running example, the eight steps are easily applied to many other business activities. For example, I have used this core methodology in areas including residential mortgage valuation, automobile collision pure premium estimation and insurance underwriting.

To some degree I have been intrigued by the lack of use of experimental design in the financial service sector, despite its obvious tremendous potential impact. But, of course, many other sectors of our economy also fail to use experimental design, despite the potential value. More interesting, however, is how isolated each of the eight business activities are from each other in most businesses. The data warehousing team, by almost universally fixing on a large and inflexible data model, locks the business into such a limited mode of operation that the data warehouse ends up ignored by the operational teams. Or, more to the current point, the data analysts work on data which simply does not contain the information the business actually needs: the impact of every controllable factor on every individual.

This isolation of activities which should be integrated is a generally observed phenomenon. While certainly there are sociological explanations for the behavior, I suggest that proactive leadership in the integration of the business processes is the only way to actually find, and build, the required synergies.

I appreciate the work of Hand et al. in further developing our thinking in special areas of data analysis and encourage them and all readers of this journal to expand our influence, both to increase the sophistication of all the tools used in our research and to integrate all the core business activities.

Rejoinder:

David J. Hand, Gordon Blunt, Mark G. Kelly and Niall M. Adams

We thank Dr. Kahn for his interesting comments.

We agree that the profit so far generated by data mining activities "has turned out to be surprisingly limited in many key businesses." We suggest that one reason for this is the poor quality of much of the data, so that the unusual pattern, when it is found, is more often due to the process of data collection than it is to the substantive content to which it relates. However, we hope that recognition and appreciation of this fact will mean that more care is put on data collection in the future, not necessarily in the way that Kahn describes (or perhaps in addition to this), but simply to ensure better quality data which may be more effectively mined.

We like the distinction between strategic and tactical business decisions. The thesis that tactical decisions should be tracked and managed so as to be continuously improved is surely one with which no one would disagree. Kahn's analysis of the continuous business learning cycle into eight operations is certainly one legitimate partition of the process. Paying careful attention to these stages, and improving those that can be improved, will doubtless lead to more effective business decisions. However, as he says, in our paper we were concerned with information-thin, data-rich problems, and it was tools for making the best of this situation that we addressed.

Surely, the point made under the heading Data Archiving, that hundreds of managers receive hundred-page weekly tallies, demonstrates the importance of the model-building aspect of data mining. Business decisions will only be aided if those "hundred-page weekly tallies" are reduced to manageable sizes.

The comments about specific business metrics under the Financial Analysis heading rang bells. We have seen many situations in which the left hand of a company did not know what the right hand was doing. Figure 11(d) in our paper might be an example of this: the exciting discovery of a dramatic change in customer behavior half-way through the time period loses its interest when one is told that, at that time, all customers were required to repay by a particular method. This is not the only example we have seen in which changes to the customer population, discovered by a risk assessment group, corresponded to shifts in marketing strategy imposed by the marketing department. This sort of thing is a powerful argument for more global perspectives on company management (and if data mining produces this result, then that alone would be a valuable consequence of the exercise).

Admittedly, the drive behind the development of data mining techniques has come from commercial sources (although the discipline is not restricted to such problems). In view of this, it will come as no surprise to hear that we entirely endorse Kahn's comment that "it is crucial to build an integrated financial model of customer activity." This is precisely the sort of model we illustrated in Hand, McConway and Stanghellini (1997) and Stanghellini, McConway and Hand (1999). On the other hand, we feel that Kahn may be a little unrealistic when he argues, under Brainstorming, that "it is thus crucial to have a view of everything a business could do. This involves explicitly knowing everything the company is trying, and has tried, both recently and long ago. Furthermore, knowledge of everything competitors are doing is obviously key." In general, while we should clearly strive to improve the information on which we base our decisions, it will inevitably be at best partial.

The comments under Experimental Design and Implementation also ring true. In various contexts, we have had considerable difficulty convincing business of the merits of accepting a small sample of high-risk customers which, while perhaps not profitable in themselves, would lead to improved decisions overall.

The example of population drift in a mail order PC market is a nice one.

We agree with Kahn that there is no doubt that formal experimental design has tremendous potential for positive impact on the financial services community. The power of such techniques is something all statisticians would presumably promote. Our paper does not contradict that. We merely argue for a parallel, in-depth investigation of the data that have been collected, and that will be collected.

Dr. Kahn's comments apply only to particular kinds of data mining, those where the aim is to optimize some response function (profit, perhaps). In other situations, however, the aim is simply to explore the data in a search for interesting structures.

ADDITIONAL REFERENCES

Deming, W. E. (1943). *Statistical Adjustment of Data*. Wiley, NY. (Republished in 1964 by Dover, New York.)

Hand, D. J., McConway, K. J. and Stanghellini, E. (1997). Graphical models of applicants for credit. *IMA J. Math. Appl. Bus. Indust.* **8** 143–155.

New York Times, 6 October 1999, page 4.

New York Times, 18 November 1999, page 4.

Stanghellini, E., McConway, K. J. and Hand, D. J. (1999). A chain graph for applicants for bank credit. *J. Roy. Statist. Soc. Ser. C* **48** 239–251.