EFFECT OF DIRTY DATA ON ANALYSIS RESULTS

(Research Paper)

Dominique Haughton
Bentley College
dhaughton@bentley.edu

Mary Ann Robbert
Bentley College
mrobbert@bentley.edu

Linda P. Senne
Bentley College
lsenne@bentley.edu

Vismay Gada
gada_vism@bentley.edu

Abstract: Information quality assessment is the process of inspecting business information to ensure that it meets the needs of the knowledge workers who depend on it. We suggest in this paper that, prior to implementing a system to assess quality, those responsible for information quality can use a subset of clean data to create a statistical model of a decision that relies on the information. Simulated perturbations of the clean data can then be used to establish a boundary for determining what degree of error produces erratic, unusable results. This approach has the advantage that it can be used to show the effects of poor data quality on the result of the analysis as accuracy declines for any reason. We do not focus on the reasons why the quality declines but rather show the consequences of poor quality data on the results of the analysis. Moreover, we examine a general error structure, one that is common in situations where errors are not additive, and is more general than that previously considered in the literature,

Key Words: Data quality, dirty data, perturbation, data quality assessment

1. INTRODUCTION

In this paper, we examine two databases: one containing data on the outcome of tax court cases; the other, loan application data. For each of these databases we build a statistical model for the outcome. Then we perturb the data to investigate how data errors propagate to affect the results of the analysis. Questions of interest in our analyses include "Should we call an expert witness?" "Should we grant the loan?" We try to answer the question, "How much error can we get away with before the analysis breaks down?" This problem falls within the class of measurement error or error-in-variables problems. [1,9.12,13]

2. PREVIOUS RELATED WORK

Both researchers and practitioners have addressed the issue of data quality. Richard Wang, Yang Lee and their associates have defined a method for assessing information quality [14]; they have also identified a set of dimensions on which to measure information quality [15]. We have chosen in this article to focus on the single dimension of accuracy. The assessment methodology as defined in [7,10] evolved from work with data from existing companies. We used this as a reference for the second data set and the first data simulation.

A number of authors have abstracted the problem to a set of equations in order to examine various aspects of the data quality problem. Haebach and Bowles defined differential equations for accuracy in terms of initial accuracy, accuracy of data being inserted, accuracy of data being updated, and proportion of data updated per annum where the last three parameters are constant [3]. This model can be used to calculate the time lag between initiation of the improvement and the eventual effect.

Lachenbruch and Mickey [6] compare techniques used to estimate error ratio rates of sample discriminate functions. They divide the empirical methods into two classes—those using a sample to evaluate a given discriminate function and those using properties of the normal distribution. They conclude that no one method is best for all solutions [6].

From a system perspective, errors noted in a simulation can be from a single database or from conflicting, unmanaged, and redundant databases containing overlapping data. In fact, Larry English defines a "schizophrenic quotient" to identify the number of redundant databases that are not controlled by replication and may have inconsistencies [2].

The results of these perturbations lead to sets of data for each of the fields where errors are probable. For each set of data there is a tuple of n values which could be compared as a set of n-tuples using a Maxi-Min methodology such as that suggested by Yager for decision making under uncertainty with ordinal information [16]. In the case of the German bank data, real options analysis as suggested by Miller [8] can be used for decision analysis to determine the effect of data quality as a factor in the uncertainties calculated.

Sometimes it is desirable to change data in a database without introducing errors that alter the characteristics of the data. Researchers have used perturbation techniques to modify data, for example, to protect sensitive information in databases. Sarathy et al [11] have examined methods for masking data in order to prevent data spies from uncovering sensitive information. However, if the data is to be useful to legitimate analysts, the perturbation methods must not introduce errors. However, our methodology does the opposite: we perturb the data specifically to introduce errors in order to study the effects of dirty data on analysis results.

3. DATABASE BACKGROUND

We have studied the effects of information quality on two distinct data sets. The first database we studied, the database on tax court outcomes, was randomly generated by the authors on the basis of typical data from publicly available transcripts, in text form, of actual cases. We also investigate a more typical second database that contained data commonly collected in business by firms and institutions authorizing credit and granting loans.

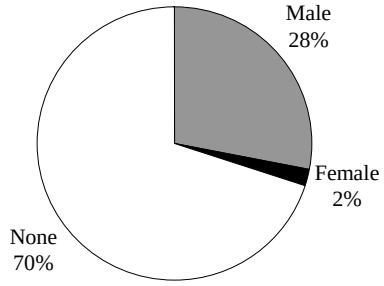
For the first part of our study, we simulated a database based on information in the court records, generating clean data by assuming a "true model" represented by a "true equation." We use simulated

data to show the effects of dirty data in a context where the outcome of the analysis is a predictive model and interest is focused on the effect of predictors such as income and gender on the dependent variable in the model. For example, loan granting institutions might be concerned with gender (or other) discrimination and thus focus on gender (or other) effects in a model which would predict who gets a loan and who does not.

On the basis of our work with court transcripts, we used a reasonable set of values in our simulations for the percentage of either a male or a female expert witness, average yearly income and the distribution of cases across regions. These factors affect the three possible outcomes of a case, i.e., win, settle or lose. The database contained 10,000 records; each record in our database has only eight fields. Note that the region field was expanded to five fields for use as predictor variables.

Expert Witness	Gender	Defendant's Income	Region1	Region2	Region3	Region4	Region5
0 if none, 1 if present	0 if male, 1 if female	$ amount	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1

Although the location of the court and the income of the defendant affected the outcome, we focused primarily on the significance of an expert witness. Testifying male witnesses dominated the 30% of the cases in which experts were called (Chart 1). The presence of an expert witness, whether male or female, significantly affected the result of the litigation (Chart 2). Defendants either won significantly more cases or suffered very few losses if an expert witness testified on their behalf. Litigation that settled fell into a band of roughly 30% to 40% of the cases, although the highest percentage of settled cases did not involve an expert.



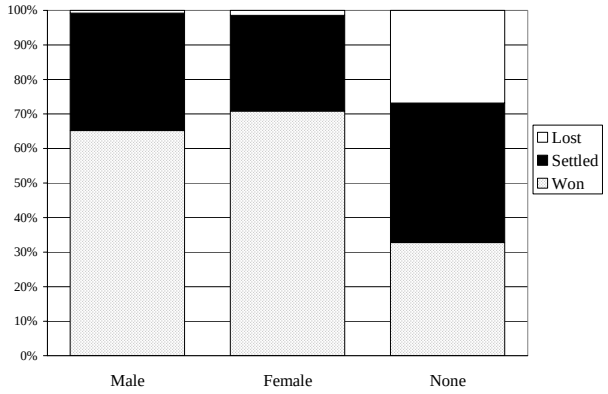


Chart 1(left): Use of Expert Witnesses in Tax Court Cases
Chart 2(right): Tax Court Case Outcomes by Gender or Presence of Expert Witnesses

This kind of database can be used to determine whether using an expert in a court of law matters. However, the reliability of this database, given that it was developed from transcripts, must be verified before it can be used dependably by lawyers deciding upon a strategy for pursuing a case. The actual database includes information on kinds of professional experience, specific courts, demographic characteristics, information technology and tools that can be used, for example, to determine whether to choose a practicing professional or an academic as an expert witness in a specific court or for a particular task. In this paper, we selected for analysis only the most significant fields to test the effect of decreased data quality.

Our second database contains data commonly collected in business by firms and institutions authorizing credit and granting loans. The loan application database we use consists of data collected by a German

bank for processing loans and has been widely used as a benchmark dataset for various data mining tools (http://www.liacc.up.pt/ML/statlog/datasets/german/german.doc.html). Each record contains 21 fields (see Appendix 1), and the database consists of 1,000 records. In order to study the effect of data quality on granting loans, we grouped fields into three categories based on the likelihood that input errors here might be a source of dirty data:

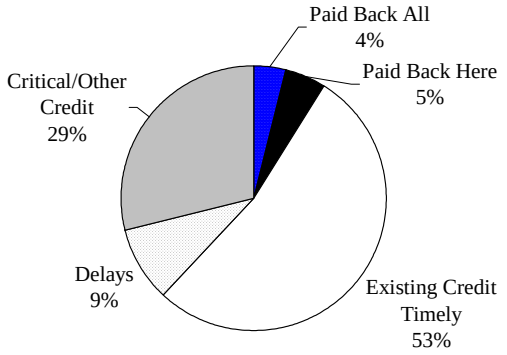
- probably clean, e.g., purpose (of the loan) and credit amount
- possibly contains errors caused by data entry, e.g., the balance in saving accounts/bonds and present employment since
- probably contains many errors since the answer to the question may not be known, the question itself is ambiguous or the information may be difficult to collect. For example, although the installment rate in percentage of disposable income is a clearly defined ratio, the data entered into the application is likely to be no more than a rough estimate. On the other hand, the options under personal status and sex are confusing:
  - Male: divorced/separated
  - Female: divorced/separated/married
  - Male: Single
  - Male: Married/widowed
  - Female: single

The analysis included here focuses on Credit history, since it is an example of data that is notoriously inaccurate in most financial databases. If there are several firms tracking an individual's credit, there is a high likelihood of inaccuracies in the data and inconsistencies among the firms.

In the German loan database, credit history is a qualitative variable that can take on one of five values:

- A30 (no credit history or all credits have been duly paid back)
- A31 (all credits at this bank have been paid back)
- A32 (existing credit is being duly repaid)
- A33 (delays in paying off in the past)
- A34 (critical account/other credits existing not at this bank)

Chart 3 shows the breakdown of the entries in this field Credit History. More than half of those applying for loans have existing credit that they are repaying in a timely fashion; about 68% of these borrowers are classified as good credit risks (Chart 4). Surprisingly, this group of borrowers is just slightly more likely to be classified as good credit than those borrowers who have had payment delays in the past.



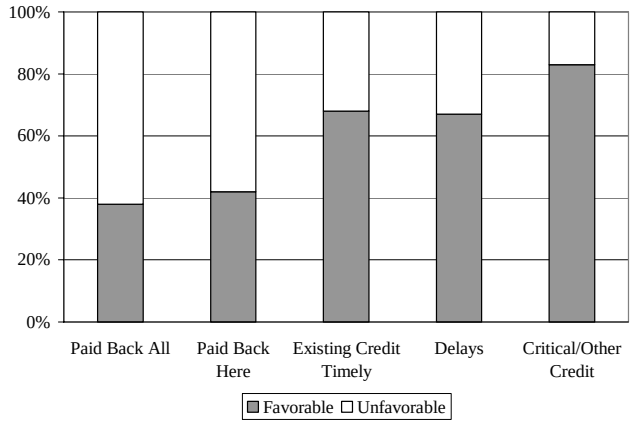


Chart 3(left): Breakdown of Data Contained in Credit History Field
Chart 4(right): "Good Credit" Classification by Credit History

The proportion of applicants classified as "good credit" in the group labeled "critical/other credit" is at first sight surprisingly high. It is likely that this group in fact consists of many applicants on the "other credit" category, who could in fact be in good credit standing for loans at other institutions. Given that these relationships exist with the original "clean" set of data, we examined the effect on these relationships when the data is dirtied proportionally as indicated by our three category levels. We examine the effects of data quality deteriorating non-uniformly across the attributes on the outcomes. Specifically, we dirtied the data by moving percentages of each field to the wrong field.

- A30 (no credit history or all credits have been duly paid back): percentage changed to A31
- A31 (all credits at this bank have been paid back): percentage changed to A30 and A31
- A32 (existing credit is being duly repaid): percentage changed to A31 and A33
- A33 (delays in paying off in the past): percentage changed to A32 and A34
- A34 (critical account/other credits existing not at this bank): percentage changed to A33

4. OVERVIEW OF OUR STUDY

In our research we discovered that the actual error structure in many situations is different from that considered in the measurement error literature. In this literature, the measurement error is typically the result of additive errors. De Varo and Lacker, for instance, use a salary discrimination example [1]. The following equations, based on their work, show earnings being determined by qualifications (QUAL) and gender (GENDER) plus some error or noise. Note that in each case the error is added to the equation.

$$y = \beta QUAL^* + \alpha GENDER + \upsilon$$

$$QUAL^* = a_0 + b_0 GENDER + u$$

$$QUAL = QUAL^* + e,$$

where υ, u, e are independent random variables with zero mean and variances $\sigma_\upsilon^2, \sigma_u^2, \sigma_e^2$; QUAL* is the true qualification of a person (and depends on GENDER as specified in the second equation); and QUAL is the measured qualification of a person.

Since in reality errors do not always follow an additive structure, we have examined a more general error structure. In our tax database, we began with our clean data and then perturbed it as follows:

- Tax-court litigation
  - gender: In our case, we simulate gender errors for a proportion of the observations so that a proportion of the ones in a gender variable is lost (gender variables contain only 1 and 0). Errors outside the 0, 1 domain can easily be eliminated, so are not considered here.
  - income: We also simulate errors which could arise from an offset error in the continuous variable, income. We multiply this variable by 10, for example, in 5% of cases and divide the resulting set by 10 in another 5% so that errors are not additive.
- Loan Data: Credit History:
  We begin to perturb the data by randomly selecting 5% of all the entries in the field credit history. We then alter the contents of this field depending on the value of the data that we had selected at random. For example, if the value in the field is A30 (no credit history or all credits have been duly paid back), we change it to A31 (all credits at this bank have been paid back). Values of A31 are changed to either A30 or A32 (with equal probability), and similarly for other values.

5. METHODOLOGY—DERIVING THE TRUE EQUATION

Multinomial logistic regression is the appropriate tool for measuring the effect of each among a set of predictors on different outcomes. There are three outcomes in the tax case: win, lose or settle (won – coded 1, settled – coded 2, lost – coded 3). We generate 10,000 observations as follows:

- Income is generated as a log-normal random variable with mean 25.610 and standard deviation 20.412 (in thousand US dollars).
- Presence of an expert witness and the expert's gender: this variable is generated as a categorical variable, with a male expert in 28% of the cases, a female expert in 2% of the cases, and no expert in 70% of the cases.
- Region is generated as a categorical variable, with 30% of cases in Region 1, 20% of cases in Region 2, 10% of cases in Region 3, 10% of cases in Region 4, and 30% of cases in Region 5.

Recall that the odds of one outcome such as win relative to another outcome such as lose are defined as the probability of the first outcome divided by the probability of the second outcome—for example, probability (win)/ probability (lose).

The dependent variable in the analysis, the outcome of the case, is generated according to the following multinomial logistic regression equations, referred to as the "true equations:"

$$\log \text{odds (win/lose)} = -.7 + .02\ \text{income} + 4.1\ \text{mew} + 3.5\ \text{few} + 1.1\ \text{region1} + .9\ \text{region2} - .6\ \text{region3} - .3\ \text{region4}$$

$$\log \text{odds (settle/lose)} = .1 + .01\ \text{income} + 3.1\ \text{mew} + 2.3\ \text{few} + .6\ \text{region1} + .3\ \text{region2} - .3\ \text{region3} - .1\ \text{region4}$$

where MEW and FEW are dummy variables for a male and female expert witness, respectively. In other words, MEW is 1 if a male expert witness was involved in the case, 0 if not. REGION1 through REGION4 are dummy variables for the first four regions. For example, region1 is 1 if the case was decided in REGION 1, 0 if not; and so on. Note that REGION5 is the reference region and is represented by a value of 0 for all four region dummy variables. Finally, INCOME denotes the defendant's yearly income in thousand US dollars.

6. TAX CASE RESULTS

Based on the cases studied, we know the actual relationship between the odds of winning or settling a case relative to losing a case. Because we generated the data based on these odds, we also know the values for the coefficients of the explanatory variables INCOME, MEW, FEW and REGION1 through REGION4. In Chart 5, the points above the label clean represent the values of the coefficients for MEW and FEW in a multinomial logistic model for win/lose and settle/lose estimated on the basis of our clean generated data. One can readily see that these values, namely 4.4, 3.9, 3.4 and 2.7, are close to the true values 4.1, 3.5, 3.1 and 2.3, although, of course, they are not identical. We now focus on how this series of altogether four coefficients for MEW and FEW vary when errors are introduced in the data.

We simulate data errors by perturbing our data in the following manner:

1. To simulate an offset error, we multiply INCOME by 10 in 5% of the cases and divide the resulting set by 10 in another 5% of the cases.
2. We first simulate a drop of 5% of female and male expert witnesses, followed successively by perturbations of 10%, 20%, 30%, 50% and 60%. As a result, in some cases female expert witnesses may be dropped entirely from the data.

Chart 5 shows the influence of the errors on the coefficients of the estimated equation. From this kind of graph, we can determine acceptable error rates for each coefficient and identify the boundary at which no useable results are found. In this chart, we plot the values of the four coefficients for MEW (male expert witnesses) and FEW (female expert witnesses) for both equations win/lose and settle/lose for a fixed error on INCOME as described in Step #1 and increasing errors on MEW and FEW as described in Step #2. There is an anomalous kink in the curve after 50% which requires further analysis.

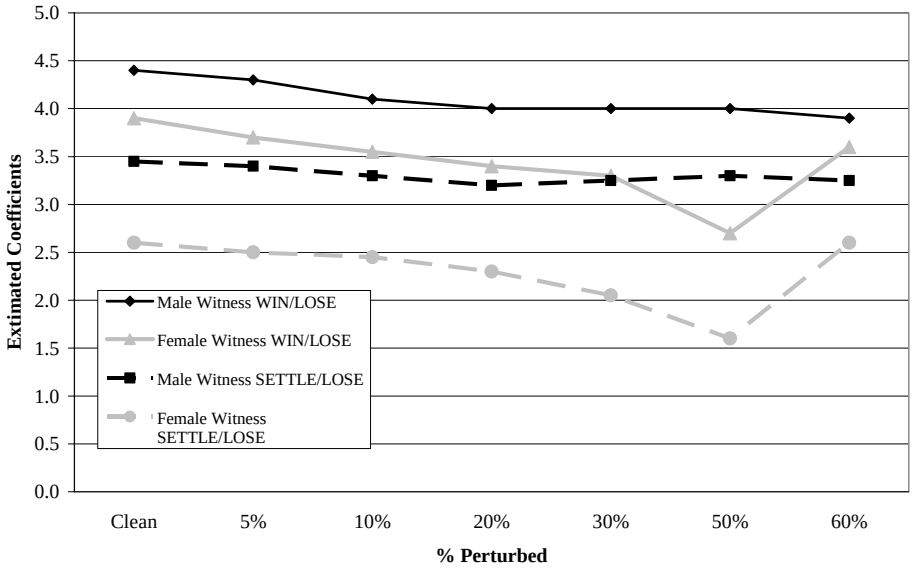


Chart 5: Estimated Coefficients with Increasing Error Rates

It is worth noting that little deterioration in the model occurs when male expert witnesses are dropped from the data in this range. However, after the 30 percent perturbation, so many female expert witnesses—a small percentage of the data to begin with—are lost that the results become unstable. The coefficients do tend to decrease with increasing error rates as observed in, for example, DeVaro and Lacker [1] where the authors found that the size of the discrimination effect—in their study, against minority loan applicants—tended to decrease with the increasing amount of data error.

At the point when 70% of the ones in the FEW variable are changed to 0's, the multinomial logistic regression model breaks down because of complete separation caused by the FEW variable: no case where a female expert witness was involved had an outcome of LOST, resulting in one of the outcome categories being empty. The following message appearing in SPSS (and in all leading statistical packages) indicates the existence of the problem:

Warnings

There is possibly a quasi-complete separation in the data. Either the maximum likelihood estimates do not exist or some parameter estimates are infinite.

The NOMREG procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

Figure 1: SPSS Separation Warning

We can see in the output given in Figure 2 that the coefficients for FEW have become very large and are in fact unreliable. This numerical phenomenon is well-known in the context of logistic regression (binary or

multinomial) when "quasi complete separation" occurs. "Separation" means that one predictor variable exactly splits the categories of the outcome variable. As we can see from the cross tabulations of OUTCOME with Female Witnesses (FEW) for error rates on FEW from 0% to 70% of the female witnesses are lost, no case is ever lost that involved a female expert witness, a fact that leads to complete separation. The cross tabulations also show that the clean data includes three observations where a case was lost that involved a female expert witness. The multinomial logistic model in fact breaks down at a point when between 30% and 50% of female expert witnesses have been lost. When 50% of few have been lost, the results are likely to be quite unreliable. Yet, for 50% of few lost, there is no SPSS warning. Consequently, there is a significant risk that the fact that the model has broken down will remain undetected.

Parameter Estimates

OUTCOME		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1 WIN/ LOSE	Intercept	.237	.053	20.208	1	.000			
	MEWPTB7	3.744	.382	96.028	1	.000	42.259	19.986	89.356
	FEWPTB7	19.804	.263	5664.076	1	.000	4.0E+08	238059181.2	667808039.6
	REGION1	.997	.077	166.430	1	.000	2.710	2.329	3.153
	REGION2	.695	.082	71.174	1	.000	2.004	1.705	2.356
	REGION3	-.552	.099	30.919	1	.000	.576	.474	.700
	REGION4	-.123	.097	1.585	1	.208	.885	.731	1.071
	INCKPTB	.003	.001	20.926	1	.000	1.003	1.001	1.004
2 SETTLE/ LOSE	Intercept	.386	.052	55.795	1	.000			
	MEWPTB7	3.079	.384	64.307	1	.000	21.736	10.241	46.132
	FEWPTB7	18.971	.000	.	1	.	1.7E+08	173303493.4	173303493.4
	REGION1	.609	.078	60.989	1	.000	1.838	1.577	2.141
	REGION2	.307	.084	13.479	1	.000	1.360	1.154	1.602
	REGION3	-.166	.091	3.342	1	.068	.847	.709	1.012
	REGION4	.039	.093	.179	1	.672	1.040	.867	1.248
	INCKPTB	.002	.001	9.725	1	.002	1.002	1.001	1.003

Figure 2: Parameter Estimates from Regression Analysis

Finally, in Chart 6, we investigate the effect of increasing error rates on INCOME on the results of the model. Chart 6 gives the coefficient of INCOME in both the estimated win/lose and settle/lose equations. There is very clear deterioration of the coefficients when INCOME is multiplied by 10 in 5% of cases and in another 5% divided by 10. Interestingly, this deterioration remains essentially the same when higher and higher proportions of cases have erroneous values of INCOME. These results indicate that the error multiplier of the actual value is more important than the percent of observations in error.

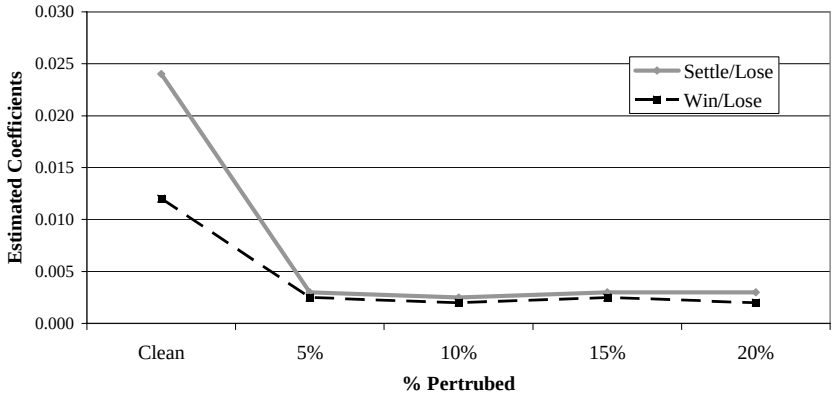


Chart 6: Estimated Coefficients for INCOME with Increasing Error Rates

7. UNDERSTANDING THE RESULTS

Given that we know that the coefficients of the variables in the regression change from the "clean coefficients" (estimated from clean data) when there are dirty data in the database, we developed a set of graphs to help us understand the impact of poor data quality on the predicted outcome of the court case. We asked, "What are the changes in the estimated probability that the defendant will win, settle or lose?" when dirty data are introduced into the database.

The charts we constructed have the probability of a given outcome—in this case, winning — estimated from dirty data (vertical axis), and from clean data (horizontal axis). If we were to plot the probability of winning based on clean data versus the same probability of winning based on clean data, the chart would contain a straight line at a 45-degree angle.

The extent to which the graph "dirty versus clean" deviates from the 45-degree line indicates the effect on the estimated probabilities of dirtying the data. If we plot the probability of winning if the data are dirty versus the probability of winning if the data are clean — as we did in Chart 7 with a 10% perturbation on the y axis—then we can clearly see that the estimated probability of winning based on dirty data is in some cases above and in some cases below that based on clean data. In other words, if we are using dirty data we will often underestimate or overestimate the probability of winning. For example, for a "clean" estimated probability of winning of .4 (on the horizontal axis), "dirty" estimated probabilities of winning range from about .2 to .5 (on the vertical axis). The data displayed in Chart 7 reflects the estimated outcome of the case in which the data for both income and expert witnesses, male as well as female, are dirty.

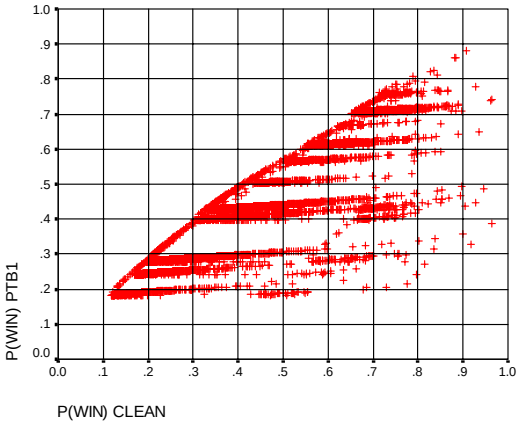


Chart 7: Estimated Probability of Winning a Case Given Dirty Data for Income and Expert Witness (Data Perturbed 10%)

While the chart suggests an overestimated or underestimated probability of winning based on dirty rather than clean data, it is not yet clear from the graph how the degradation of each factor—income or the presence of an expert witness—separately affects the shift in the estimated probability of each outcome.

To gain insight into these separate effects, we perturb the data on expert witnesses leaving the data on income (and regions) clean. Chart 8 on the following page shows the deterioration in the estimated probability of winning with just 10 percent of the expert witness data dirty. When we plot the estimated probability of settling with inaccurate information on expert witnesses (versus that for clean data), the probability is overestimated for low probabilities but underestimated at higher probabilities (Chart 9). And finally, if the expert witness data is wrong, at very low probability ranges (near zero on the horizontal axis), the estimated probability of losing is overestimated, dramatically in some cases

Chart 8: Pro
Data for Expe

Chart 9: Prob. of Setting a Case Given Dirty Data for Expert Witness (10%)

Chart 10: Prob. of Losing a Case Given Dirty Data for Expert Witness (10%)

Chart 11: P
for Expert W

a

(Chart 10). By viewing the charts, we can study the patterns of changes in estimated outcome caused by the deteriorating quality of a given variable. For example, Charts 11, 12 and 13 (where 60% of the expert witness data was perturbed but data on income were kept clean) indicate a pattern similar to that in Charts 8, 9 and 10 but with a moderately more severe tendency to overestimate or underestimate the probability of each outcome.

By comparing Charts 7 and 8 we can see that the horizontal streaks in Chart 7 as well as the shift of its diagonal boundary away from the 45% line are due to perturbations of the INCOME variable.

8. GERMAN CREDIT DATA RESULTS

We also use binomial logistic regression for measuring the effect of each among a set of predictors on the two different outcomes for a loan application—granted or denied—based on the 1,000 observations in the German database.

Log odds (yes/no) = +4.41 -1.55 overdrawn -1.17 0-to-200dm
-1.73 paid-back-all - 1.65 paid-back-here - .71 paid-back-previously
- .67 delays - .75 new-car + .61 used-car - 1.05 repairs
-.87 below100dm_savings - .60 100-to-500dm_savings - .50 single-male
-.04 duration -.19 installment rate

We then follow the process of perturbing the loan data described earlier. Chart 14 reveals that the deterioration in the coefficients for the credit history dummy variables is moderate for perturbations of these variables on up to 20% of the database but then the anomalous behavior pattern that occurred in the Witness Data appears. The coefficients are clearly not stable.

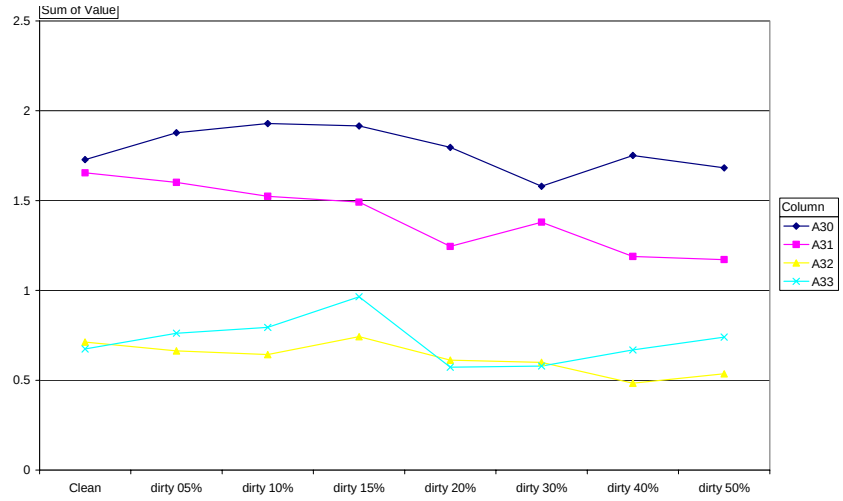


Chart 14: Estimated Coefficients with Increasing Error Rates

When the data for the INSTALLMENT RATE is perturbed, the regression analysis again produces anomalous results. Chart 15 shows the results when there are multiple effects of dirty data. For example, the installment rate data for loan applicants itself can be incorrect. This fact is shown in the chart by the first number on the x axis, e.g., "d05%" means that 5% of the data is dirty. In addition, the interest rate used in the analysis is also dirty, e.g., + - 05% means the rate itself is off

by plus or minus 5%. The coefficients did not deteriorate steadily but rather oscillated before falling off.

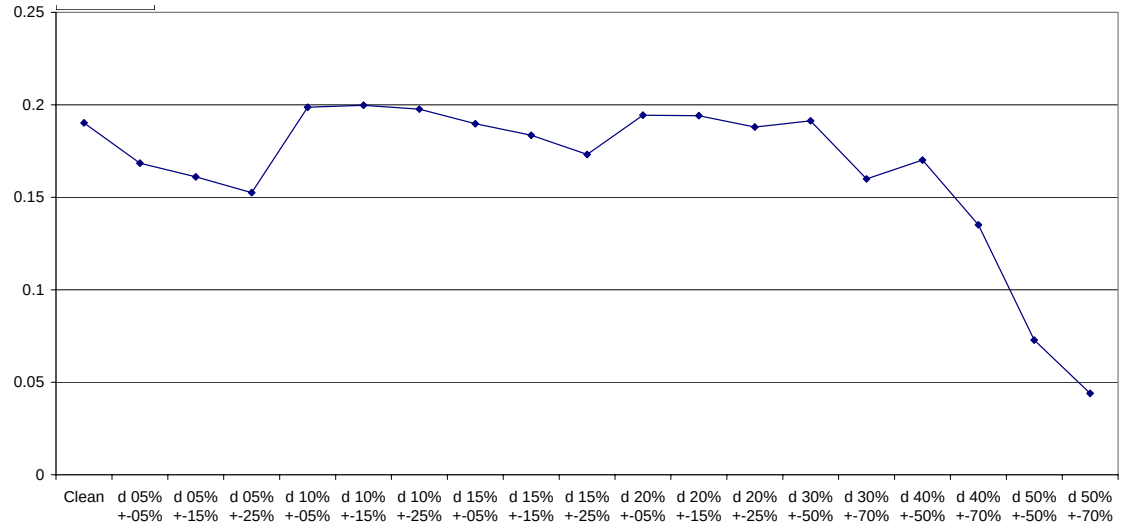


Chart 15: Both Interest Rate Data and Interest Rate Both Perturbed

9. CONCLUSION

One hundred percent data accuracy is not required for all projects; in fact, it might not even be possible. We posed the question, "What is the effect of the error rate on the results produced?"

In this paper we have analyzed the effect of decreasing data quality on the results of logistic and multinomial logistic regression results. By determining critical points in the acceptability of errors, we can establish boundary points and ranges where the derived results can no longer be trusted.

Our simulations provide a roadmap that could be followed in situations where data quality is at issue and the objective of the analysis is a predictive model, as is the case in direct marketing or in investigations of possible discrimination. The process essentially involves building a predictive model with clean data and measuring the effects on the model of progressive simulated deteriorations of the input data. Our results suggest that the effects of dirty data are not uniform and predictable. For example, the model of the tax-court outcome deteriorates gradually as the input data deteriorates until the model reaches a boundary point—identified in our results—where the model becomes quite unreliable.

Our results to date indicate that perturbations of a categorical variable on even a relatively high proportion of the database are relatively less damaging to analysis results than offset errors on numerical variables (comparing the deterioration of results due to perturbed income in the tax-court case to that due to perturbed gender of expert witnesses). Ideally, a sample of clean data is available to build a clean model to use as a benchmark against which to compare deteriorated models. In the absence of such a clean sample, sensible values (derived from the nature of the problem at hand) are needed to randomly generate clean data as we did in the expert witness case.

The decision on how to perturb the data or what degree of quality is required for a specific application is a separate study. Though models selected will not be a perfect analogy, methods of evaluation such as impact analysis [5] or real options analysis [8] will assist in the selection of the best alternatives. Such a study of decision making under uncertainty has not been attempted in this paper but is a logical extension of our work.

Since the degradation in the regression results is erratic rather than linear, decisions based on the results of the analysis could be incorrect. We intend to examine larger data sets to determine the pattern of oscillation. Specifically, we will extend our investigation to other situations where the objective of the analysis is whether to grant a loan. We plan to track the propagation of errors in the input data to the probability of making one decision versus another. Our future work will involve more general error structures and a possible perturbation of the dependent variable.

REFERENCES

[1] De Varo, J. and Lacker, J.,(1995). Errors in Variables and Lending Discrimination. *Federal Reserve Bank of Richmond Economic Quartely*, Vol. 81, No.3.

[2] English, L., (1999) *Improving Data Warehouse and Business Information Quality*. John Wiley & Sons, New York, NY.

[3] Haebich, W., Bowles, S & Associates, (1997). A Quantitative Method to Support Data Quality Improvement. *Proceedings of the 1998 Conference on Information Quality*.

[4] Huang, K., Lee, Y. and Wang, R., (1999). *Quality Information and Knowledge.* Prentice Hall, Upper Saddle River: NJ.

[5] Ketchel. J. and Dolan, J. (1974). Impact Analysis. *Proceedings of the 1974 ACM Annual Conference.* 318-325.

[6] Lachenbruch, P. and Mickey, M., (1968, February). Estimation of Error Rates in Discriminant Analysis. *Technometrics*, Vol. 10, No. 1, 1-11.

[7] Lee, Y., Strong, D., Kahn, B. and Wang, R., (accepted 2001, November).AIMQ: A Methodology for Information Quality Assessment. *Information and Management.*

[8] Miller, L. and Park, C. (2002). Decision Making Under Uncertainty-Real Options to the Rescue. *The Engineering Economist.* Vol. 47, No.2, 105-150.

[9] Naik, P. and Tsai, C.,(2000). Controlling Measurement Errors in Models of Advertising Competition. *Journal of Marketing Research*, Vol. 37, 113-124.

[10] Pipino. L., Yang, L. and Wang, R.,(2002, April). Data Quality Assessment. Communications of the ACM. Vol. 45, No. 4, 211 – 218.

[11] Sarathy, R., Muralidhar, K., and Parsa, R., (2002). Perturbing Nonnormal Confidential Attributes: The Copula Approach. *Management Science*, Vol. 48, No.12, 1613-1627.

[12] Shilling,J., (1993). Measurement Error in FRC/NCREIF Returns on Real Estate. *Southern Economic Journal*, Vol. 60, No.1, 210-219.

[13] Stanley. T., (1988). Forecasting From Fallible Data: Correcting Prediction Bias With Stein-Rule Least Squares. *Journal of Forecasting*, Vol. 7, 103-113.

[14] Strong,D., L. and Wang,R., (1997, May) Data Quality in Context. *Communications of the ACM*, Vol. 40, No 5,103-110.

[15] Wand, Y and Wang, R, (1996, November). Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, Vol. 39, No.11, 86-95.

[16] Yager, R. (1999) Decision Making Under Uncertainty with Ordinal Information. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 7 No5, 483-500.

APPENDIX

Description of the Variables in the German Credit Data

Original source of data:

Professor Dr. Hans Hofmann
Institut fur Statistik und Ökonometrie
Universität Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

	Attribute	Possible Values	Meaning
1	Qualitative Status of existing checking account	A11	< 0 DM
		A12	0 to < 200 DM
		A13	Salary assignments for at least 1 year
		A14	No checking account
2	Numerical Duration in month		
3	Qualitative Credit history	A30	No credits taken/all credits paid back duly
		A31	All credits at this bank paid back duly
		A32	Existing credits paid back duly till now
		A33	Delay in paying off in the past
		A34	Critical account/other credits existing (not at this bank)
4	Qualitative Purpose	A40	Car (new)
		A41	Car (used)
		A42	Furniture/equipment
		A43	Radio/television
		A44	Domestic appliances
		A45	Repairs
		A46	Education
		A47	(vacation - does not exist?)
		A48	Retraining
		A49	Business
		A410	Others
5	Numerical Credit amount		
6	Qualitative Savings	A61	< 100 DM
		A62	>= 100 DM but < 500 DM

	Attribute	Possible Values	Meaning
	account/bonds	A63	<= 500 DM but <1,000 DM
		A64	>=1,000 DM
		A65	Unknown/ no savings account
7	Qualitative	A71	Unemployed
	Present employment	A72	< 1 year
	since	A73	>=1 but < 4 years
		A74	<= 4 but < 7 years
		A75	>= 7 years
8	Numerical		
	Installment rate in % of disposable income		
9	Qualitative	A91	male: divorced/separated
	Personal status and		
	sex	A92	female: divorced/separated/married
		A93	male: single
		A94	male: married/widowed
		A95	female: single
10	Qualitative	A101	
	Other	A102	
	debtors/guarantors	A103	
11	Numerical		
	Present residence		
	since		
12	Qualitative	A121	
			If not A121: building society savings
	Property	A122	agreement/life insurance
		A123	If not A121/A122: car or other not in attribute 6
		A124	unknown/no property
13	Numerical		
	Age in years		
14	Qualitative	A141	
	Other installment		
	plans	A142	
		A143	
15	Qualitative	A151	
	Housing	A152	
		A153	
16	Numerical		
	Number of existing credits at this bank		
17	Qualitative	A171	unemployed/unskilled - non-resident
		A172	unskilled - resident
		A173	skilled employee/official
			management/self-employed/highly qualified
		A174	employee/officer
18	Numerical		
	Number of people being liable to provide maintenance for		
19	Qualitative		
			yes, registered under the customer name
20	Qualitative	A201	Yes

	Attribute	Possible Values	Meaning
	Foreign worker	A202	No
21	Outcome:	1	Good credit
		2	Bad Credit