The effects and interactions of data quality and problem complexity on data mining

(Research Paper)

Roger H. Blake
University of Massachusetts Boston
roger.blake@umb.edu

Paul Mangiameli
University of Rhode Island
mangia@uri.edu

Abstract: Data quality remains a persistent problem in practice and a challenge for research. In this study we focus on four of the most important dimensions of data quality - accuracy, completeness, consistency, and timeliness. Definitions and conceptual models for these dimensions have not been collectively considered with respect to data mining in general and a key determinant of data mining outcomes, problem complexity, in particular. Conversely, these four dimensions of data quality have only been indirectly addressed by data mining research. Using definitions and constructs of data quality dimensions, our research shows for the first time that data quality and problem complexity have significant interaction effects on data mining outcomes. It also shows that the dimension of consistency can be concluded to have the largest effects of all four. Our results help address an important research challenge noted by March and Smith [15]. They also enable practitioners to assess data quality, prioritize efforts to improve it, and increase the performance of data mining for better decisions.

Key Words: IQ Concepts, Tools, Metrics, Measures, and Models, Business Intelligence and IQ, Data Warehouses and Data Mining

Introduction

There are no universally agreed upon definitions of data quality [24], or even of quality itself [9]. However, there is no dispute about the importance of data quality and the consequences it can have when poor. Examples of the impact of poor data quality are easy to find and range from minor incidents to major financial consequences [24] to incalculable losses [26].

Our study is of four dimensions of data quality: accuracy, completeness, consistency, and timeliness. This set of dimensions was chosen because they are among the most objective in Wang and Strong's framework [30] and because of their significance - the "dimensions of accuracy, completeness, consistency, and timeliness have been widely cited in the literature as the most important data quality dimensions to the information consumers" [20]. We develop a set of assessment metrics for each of the four dimensions and evaluate how using those metrics to vary levels of data quality can affect data mining outcomes.

Despite the significance of both data quality and data mining research, each remains a largely distinct stream. In data mining research, the concept of an accuracy dimension is most closely matched to studies of the effects of noise [e.g. 33] and the concept of a completeness dimension to studies of missing values [e.g. 13]. However, data mining researchers do not typically refer to data quality dimensions nor give explicit consideration to data quality constructs. On the other

hand, problem complexity is a fundamental to data mining research. Extensive research has been conducted to improve data mining performance for scenarios of higher problem complexity, for example Quinlan's work to develop ID3 [23], but problem complexity is not generally studied in data quality research.

Conversely, problem complexity is not a direct consideration in data quality research, possibly because it cannot be tied to any single data quality dimension. Increases in problem complexity might be associated with changes in data quality along one or possibly more dimensions, yet both are factors that can influence data mining outcomes.

Ge and Helfert [7] highlighted the importance of defining and assessing data (or information) quality dimensions as well as applications of those assessments. In industry the application of data mining has become increasingly important as analytical techniques have become widely recognized as a key to gain competitive advantage [27] and enable tighter integration between business process systems and decision-making [11]. Our study is one of the first to draw data quality metrics from data quality literature to measure data quality and evaluate its effects and interactions with problem complexity.

Our research demonstrates that significant interaction effects exist between data quality and problem complexity on the outcomes of data mining, and it shows the conditions under which those effects occur. These are important findings to researchers seeking to define metrics and assess data quality and to those seeking to improve data mining algorithms under varying conditions of data quality along multiple dimensions. The findings in our study can help practitioners recognize how data quality will affect their use of data mining tools and how they prioritize data quality improvement efforts.

Background

The following subsections present our definitions and metrics for each dimension of data quality, problem complexity, and data mining outcomes. A discussion of the methods to equate those metrics to measurement levels as factors in the experimental design is held until after presenting the hypotheses.

Please note that our metrics are developed and applied using a categorical class variable and a single continuous attribute variable. Problems with this structure occur frequently; typical examples include the prediction of which customers will respond to a direct marketing promotion and which credit applicants are credit-worthy. In order to isolate the effects of data quality from problem complexity, problem complexity was represented only in the class variable and data quality was represented only in the attribute variable. The following sections present our definitions and metrics for each dimension of data quality, problem complexity, and then for data mining outcomes.

Metrics for data quality assessment

A value expressed in a range from 0 to 1 has long been considered a desirable trait for a data quality metric [4]. Metrics in this form have often been used especially for the intrinsic dimensions of data quality such as we are considering. Examples include metrics for completeness and consistency in the theoretical model developed by Ballou and Pazer [3] and the metrics for completeness by Shankaranarayanan and Cai [25]. We adopt the same convention and use a range of 0 to 1 to represent lowest to highest data quality.

To assess quality we use metrics in two functional forms described by Pipino et al. [22]. These authors suggested simple ratios for the more objective dimensions of data quality. Simple ratios are typically based on percentages of data items meeting specific criteria, such as the percentages of data items which are complete. We follow suit by assessing data quality with simple ratios for the dimensions in which they are applicable: accuracy, completeness, and consistency. The fourth dimension of timeliness uses a min/max metric as described later.

Of the four dimensions we study accuracy has a particularly wide range of definitions. Many consider accuracy as meaning a correct and unambiguous correspondence with the real world. One example of this view defines accuracy as meaning "the recorded value is in conformity with the actual value" [2]. Other definitions based on this correspondence include accuracy as "agreement with either an attribute of a real world entity, a value stored in another database, or the results of an arithmetic computation" [10]. Accuracy has also been defined as a group of intrinsic dimensions such as completeness, consistency, or timeliness. A firm definition is elusive; as Wand and Wang summarized, "there is no commonly accepted definition of what it means exactly" [28]. Notwithstanding, researchers have used metrics for accuracy based on the rate of correct data items over an entire relation, using a 1 for an accurate data item, and a 0 otherwise. A metric in this form is not designed to take into account the magnitude of differences between correct and incorrect values, only whether a data item is correct or not. Wang et al. [31] measured accuracy this way for zip codes in a customer table as did Motro and Rakov [17] for a data warehousing environment. We do the same and define an accuracy metric A based on a simple ratio as:

$$A = 1 - \left(\sum_i^N f(d) / N \right)$$

where N is the number of data elements, and $f(d)$ is 0 if data element d is correct, and 1 otherwise.

Complete data has been defined as data having all values recorded [8], and data having the "presence of all defined content at both data element and data set levels" [3]. Relevant to our study is that data mining algorithms generally do not differentiate between categories of missing values, and so for our metric we adopt Shankaranarayanan and Cai's [25] definition of completeness as the ratio of the values that have been recorded to those that could possibly have been recorded without regard to cause. Our metric C_p is similar to Ge and Helfert's [6] ratio and is defined as:

$$C_p = 1 - \left(n / (N(1 + A)) \right)$$

where N is the number of instances in a relation, A the number of attributes, and n the number of data items with null values.

Definitions of consistency often refer to uniformity. Ballou and Pazer [3] defined consistency as when "the representation of the data value is the same in all cases" and as "format and definitional uniformity within and across all comparable data sets" [3]. Gomes [8] defined data as consistent "if it doesn't convey heterogeneity, neither in contents nor in form", and uniformity was again expressed in a definition of consistency as being related to ambiguity and the "same value repeatedly expressed for the same situation in the real world" [28].

The degree of uniformity can be found in consistency metrics based such as Wang et al.'s that are based on the percentages of tuples with violations of referential integrity [31]. We developed our consistency metric C_n with respect to referential integrity and their work, defining our metric as:

$$C_n = 1 - (V / N)$$

where N is the number of tuples in the relation and V the number of tuples with violations of referential integrity.

To represent inconsistency in our datasets an attribute value was switched to the value of another attribute selected at random from the instances in a different class. This switch creates an overlap between class boundaries and an ambiguity whereby instances with the same attribute values can belong to two (or more) classes. A higher rate of switched attribute values represents higher ambiguity and therefore a lower level of the metric for consistency.

Our consistency metric and the method used to create differing levels of consistency closely correspond to Ordonez and García-García's [19] consistency metric. These authors developed and explored their metric through an example of two relations, R and S. Following their example, relation S had a primary key k and attribute f. R was a de-normalized relation with foreign key k and foreign attribute f, and S was a relation referenced by R, such as can be found in many data warehouses. These author's metric was based on the number of tuples in R with non-null values of R_k in which $R_k=S_k$ and $R_f=S_f$ taken as a percentage of the number of tuples in R with non-null values of R_k in which $R_k=S_k$. In general terms this metric is based on the number of tuples with matches between R and S for both k and f, divided by the number of tuples with matches between R and S for only k. In more formal terms this metric can be expressed as $\Im_{Count(R.k)} (\sigma_{R.k\neq null} (R \bowtie_{R.k=S.k,R.f=S.f} S)) / \Im_{Count(R.k)} (\sigma_{R.k\neq null} (R \bowtie_{R.k=S.k} S))$. The datasets used in our study are the equivalent of R with class variable k and attribute variable f. In the case of perfect consistency there is a relation S for which there are no mismatches, $\Im_{Count(R.k)} (\sigma_{R.k\neq null} (R \bowtie_{R.k=S.k} S)) = \Im_{Count(R.k)}$. Changing a value R_f to correspond to another class in R_k by definition means that there will be a tuple in R which will not match on both R_k and R_f. For example, if the original value of a tuple's attribute was $R_{f=p}$ and it was changed to $R_{f=q}$, then the denominator of the metric $\Im_{Count(R.k)} (\sigma_{R.k\neq null} (R \bowtie_{R.k=S.k} S))$ would be unaffected.

However, this change would create an unmatched pair of k and f values in R and S and the numerator of the metric would decrease by 1. This mismatch would mean that $\Im_{Count(R.k)} (\sigma_{R.k\neq null, R.k=S.k, R.f=q, S.f=p}(R \bowtie S)) = 0$ and the consistency metric for a dataset decreases in value as ambiguity is added. Using these conditions, our metric for consistency would be the same as that of Ordonez and García-García metric if both were to be calculated from the same dataset.

Timeliness has been defined as inversely related to the degree data is out of date [2]. Volatility is the length of time between real world change and a subsequent change which invalidates the original data, and currency is the length of time between real world change and data input. The three component points of time used in these definitions are shown in Figure 1:

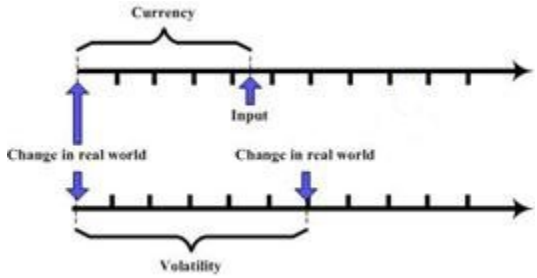

Figure 1: The points in time used to define timeliness

When multiple indicators of data quality need to be aggregated into a single metric such as those using currency and volatility, Pipino et al. [22] suggested min/max operators. We follow this form and base our metric on one defined for timeliness by Su and Jin [26] which is:

$$Max\left(0,1-\left(Currency/Volatility\right)\right)$$

To represent varying degrees of currency and volatility we first derived a classification model from a generated dataset containing training and test dataset and then injected varying percentages of new instances for re-classification by the model. We took currency to be fixed – the data in a dataset has already been inputted - but volatility to vary as differing percentages of new data replaced existing, potentially stale, data. From these assumptions our metric for timeliness T is defined as:

$$T = \left((N-R)/N\right)$$

where N is the number of instances in the training and test data and R the number of new instances introduced for re-classification.

Metric for problem complexity

Our problem complexity metric is based on the entropy of the class variable and only the class variable. Many established measures of problem complexity use entropy directly. Alternative representations such as those based on the information content of patterns found in data often produce results that are similar to, if not the equivalent of, entropy [14]. Of equal importance to our study is that many data mining algorithms for classification, including the algorithm we use, integrate entropy into the criteria for tree induction and pruning [16]. Entropy is a measure of uncertainty contained in data and defined using the probabilities of membership in each of the classes of a categorical class variable:

$$\sum_{i=1}^{n} p\left(x_i\right)\log p\left(x_i\right)$$

where there are n classes and $p(x_i)$ is the probability of any instance belonging to class i.

Using alternate means to calculate problem complexity such as conditional or mutual entropy would include the attribute variable and the class variable and introduce a potential co-variance. To avoid this confound, problem complexity was represented only in the class variable.

Metric for data mining outcomes

The F-measure is a popular metric for representing the outcomes of data mining because it can be expressed as a single figure [32]. This measure is based on a combination of recall and precision which can be found in the confusion matrix produced by evaluating a test dataset with a classification model. Recall measures the percentage of instances in a test dataset belonging to a category of a class variable that the model correctly identifies. Precision measures the percentage of instances the model identifies as belonging to a category of a class variable that are correct. The elements of a confusion matrix are shown in Table 1.

	Actual positive	**Actual negative**
Predicted positive	*TP*	*FP*
Predicted negative	*FN*	*TN*

Table 1: Elements of a confusion matrix

where:

TP = Percentage of True Positives,
FP = Percentage of False Positives,
TN = Percentage of True Negatives,
FN = Percentage of False Negatives, and

$TP + FP + TN + FN = 100\%$.

From those elements recall and precision are defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{and}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

where *TP*, *FP*, *TN*, *FN* are defined as above.

From the harmonic mean of precision and recall, the F-measure is calculated as:

$$\text{F-measure} = \left(2 \cdot \text{Precision} \cdot \text{Recall}\right) / \left(\text{Precision} + \text{Recall}\right)$$

Precision and recall are weighted equally when the F-measure is calculated in this way. An alternative calculation weights each differently to balance differences in the desirability of a false positive and a false negative. We choose equal weights but note that unequal weightings for specific applications could readily be incorporated in our experimental design.

Hypotheses

Each hypothesis posits significant effects on data mining outcomes from two factors: data quality and problem complexity. In our experimental design a continuous attribute variable is used for representing varying levels of data quality and a categorical class variable for varying levels of problem complexity. However, we do not know if our metrics for data quality and problem complexity will bear a relationship and possible correlation to one other. Separating data quality from problem complexity in two variables isolates the effects of each on data mining outcomes. To the extent that the metrics for data quality and problem complexity are related, separating them ensures that any significant effects found for data quality or problem complexity are significant when controlling for the other factor.

There are four major hypotheses, each consisting of a group of analogous individual hypotheses. Each individual hypothesis involves a single dimension of data quality and the effects of one or more levels of data quality and one or more levels of problem complexity on data mining outcomes. The null hypothesis for each is that there will be no significant effects of data quality or of problem complexity on data mining outcomes. Each is stated in the positive form as an alternate hypothesis that there are significant effects. None of the four major hypotheses are intended to be collective; rather, analogous individual hypotheses have been grouped together to avoid a lengthy list of highly similar hypotheses.

Hypothesis I: Taken individually, each level of data quality or level of problem complexity will have a significant effect on data mining outcomes as indicated by the F-measure.

Hypothesis I is comprised of five individual hypotheses, one each for accuracy, consistency, completeness, timeliness, and problem complexity, all for their effects on the outcomes of data mining.

Hypothesis II: For any given level of problem complexity, as the level of data quality in any given dimension decreases there will be a corresponding negative and significant effect on data mining outcomes as indicated by the F-measure.

Hypotheses II is concerned with the impact of the effects of data quality on data mining outcomes. To establish whether there are significant main effects, and if so the direction of those effects, comparisons between the low and medium, medium and high, and low and high levels of data quality while holding the level of problem complexity constant are required. Since three comparisons of data quality are to be made for each of three levels of data quality and four dimensions of data quality, 36 individual analogous hypotheses comprise Hypothesis II. Table 2 enumerates the nine main effects hypothesized for accuracy in Hypothesis II; an analogous set is hypothesized for the remaining three dimensions of data quality.

	Treatments		Control	Hypothesized main effects on data mining outcomes
Dimension	Data quality comparisons		Complexity	
Accuracy	High	Medium	Low	☐ Significant
Accuracy	High	Medium	Medium	☐ Significant
Accuracy	High	Medium	High	☐ Significant
Accuracy	High	Low	Low	☐ Significant
Accuracy	High	Low	Medium	☐ Significant
Accuracy	High	Low	High	☐ Significant
Accuracy	Medium	Low	Low	☐ Significant
Accuracy	Medium	Low	Medium	☐ Significant
Accuracy	Medium	Low	High	☐ Significant

Table 2: The nine comparisons evaluating the impact of decreasing levels of data quality (Hypothesis II)

In addition, Hypothesis II states that not only will there be significant effects of data quality on data mining outcomes, but that those outcomes will decrease with lower levels of data quality. A graphical representation in Figure 2 shows the posited effects in Hypothesis II to be evaluated for each comparison between two levels of data quality.

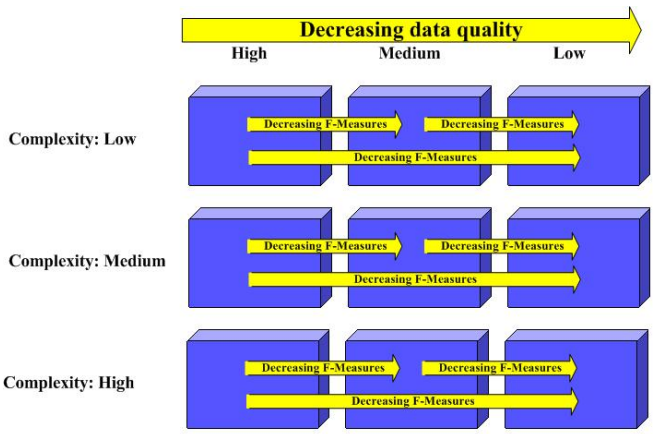

Figure 2: Hypothesized direction of effects in Hypothesis II

Hypothesis III: For any given level of data quality in a stated data quality dimension, as the level of problem complexity increases there will be a corresponding negative and significant effect on data mining outcomes as indicated by the F-measure.

Hypotheses III is concerned with the significance and direction of the main effects of problem complexity levels on data mining outcomes. To establish whether there are significant main effects, and if so the direction of those effects, there needs to be comparisons between the low and medium, medium and high, and low and high levels of problem complexity while holding the level of data quality constant. Since there are three comparisons of problem complexity to be made for each of three levels of data quality and four dimensions of data quality, 36 individual similar and analogous hypotheses in total make up Hypothesis III. Each of the comparisons involves the effects of two levels of data quality as treatments while holding a single level of problem complexity constant. Figure 3 illustrates the hypothesized direction of main effects for comparisons to be made in evaluating Hypothesis III.

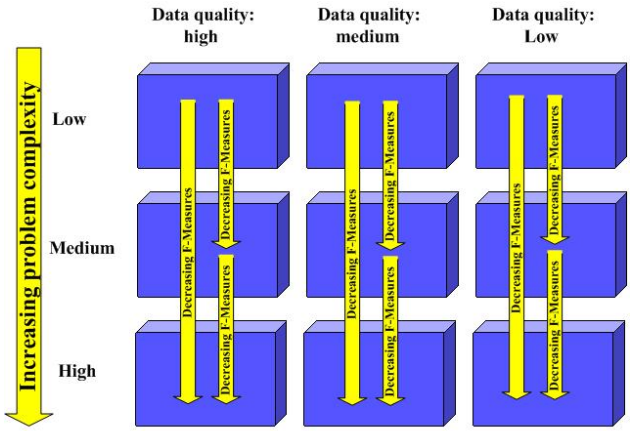

Figure 3: Hypothesized direction of effects in Hypothesis III

Hypothesis IV: Taken together, as the data quality level of any single data quality dimension decreases and the level of problem complexity increases, their interactions will have a corresponding significant and negative effect on data mining outcomes as indicated by the F-measure.

Hypothesis IV has one individual hypothesis for each dimension of data quality. Each hypothesis is based on a statement that the interaction effects of single dimensions of data quality and problem complexity on data mining outcomes will be significant. Furthermore, as data quality degrades or problem complexity increases, the individual effects of each on data mining outcomes will be negative and significant. These hypothesized effects are represented in Figure 4.

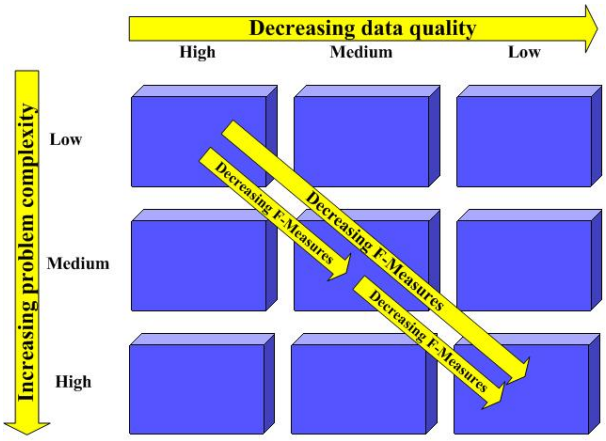

Figure 4: Direction of interaction effects hypothesized in Hypothesis IV

The main effects hypothesized in Hypotheses I through III are intended to build towards Hypothesis IV and evaluating interaction effects. The effects to be evaluated in Hypotheses I through III are all expected to be significant; findings of significant interaction effects in Hypothesis IV and any patterns in those effects would be useful in determining how efforts to improve data quality should be prioritized and how it affects data mining.

Presented next is the experimental design for evaluating Hypotheses I through IV, first for how data quality and problem complexity metrics were equated to levels to use as factors, then for the methods used to generate datasets.

Experimental design

Data quality measurements

To measure their effects, specific ranges of values for the data quality metrics as described above were related three levels of data quality. Using levels to measure data quality has been established in the literature. Some have measured data quality in two levels, low and high [29]. We use three levels of each metric to measure data quality to be consistent with Parssian's research [20]: "Often, subjective qualitative measures, such as low, medium, and high, are used to indicate the quality of data."

The metric for accuracy is defined as the rate (percentage) of correct attribute values in a dataset. In our datasets we introduce inaccuracy by randomly selecting and changing an attribute value. To avoid any inadvertent correlation of the attribute with the class that could come through systematic perturbations, attribute values were changed to that of another tuple at random.

Prior research provides several guidelines for appropriate rates of incorrect values. Parssian [21] discussed accuracy levels from three sources ranging from 80% to 90%, and others have simulated accuracy rates in the range of 90%. [5]. Klein et al. found that even the most critical databases have estimated inaccuracy rates of being from 1% to 10% [10]. Given their critical nature presumably those are databases of high quality. Based upon our preliminary investigation with similar datasets, a high level of accuracy was established as having a range of 92-100% correct values (a mean of 96%), a medium level as 88-92% (a mean of 90%), and a low level as 80-88% (a mean of 84%).

The metric for completeness is an attribute level metric derived from the rate of nulls inserted into an attribute. From base datasets generated as described above, different rates of null values were inserted into the attribute variable to represent different levels of completeness. Guidelines from past research indicate a wide range of null value rates have been used to represent completeness. Parssian et al. [21] discuss examples of completeness rates ranging from 75% to 95%. Others have experimented with 50% null attribute values [6], and some have reported databases with missing value rates of 50% and more [12]. Based on this we use the same mean rates and ranges of completeness as for accuracy.

The consistency metric by Ordonez and García-García was well developed but those authors did not explore particular values to represent levels of data quality. Referential integrity metrics developed elsewhere also provide little guidance. We chose the parameters for the three measurement levels to be consistent with accuracy and completeness for both mean values and the range of consistency.

Prior work does not suggest values of the timelines metric for low, medium, or high levels. In our datasets varying levels of timeliness were represented by varying the proportions of new instances replacing old instances in a test dataset. We ran our data mining algorithm with 10-fold cross validation to avoid any exaggeration of effects. A high level of timeliness was represented by retaining an average of 82% of the instances, a medium level by retaining 50%, and a low level by 18%.

Problem complexity measurement

As stated earlier, entropy is used as the metric for problem complexity. Entropy is related to the number of categories a class variable has – at one extreme is a class variable with one category and therefore no complexity, while towards the other extreme is a class variable many categories and high complexity. To develop measurement levels a Monte-Carlo simulations and a *k*-means cluster analysis was conducted which determined that low complexity was best represented by a class variable having two categories, medium complexity by a class variable having three or four categories, and high complexity as a class variable having five through eight categories.

Dataset generation

The data mining problems we analyzed each had a categorical class variable and one continuous attribute variable. Data with this structure corresponds to a relation extracted from a data warehouse, such as those having one identifier and one non-identifier in the schema used for data quality analysis by Parssian [1]. This structure also equates to the relations having a foreign attribute and a foreign key analyzed by Ordonez and García-García [19] to develop data quality metrics.

Each data mining problem classified a dataset with 300 instances. This number of instances is an approximate performance threshold beyond which improvements in the performance of

classification algorithms may not be appreciable [18]. Each of these datasets was created as a base dataset with a specific level of problem complexity and perfect data quality, and from that varying levels of data quality were induced. The attribute variable was generated from a uniform distribution.

The sequence for our analysis begin with the generation of a sample of 100 datasets for each combination of the four data quality dimensions, three data quality levels, and three problem complexity levels. Each of these 100 datasets was analyzed by the J48 classification algorithm and the F-measure was calculated. One observation for subsequent statistical analysis consisted of a single F-measure derived from one dataset and the corresponding metrics for problem complexity and data quality used to generate that dataset.

In order to make the analysis of a large number of combinations of dimensions, data quality levels, and problem complexity levels feasible, an application was written to automate the entire procedure from dataset generation through statistical analysis and hypothesis testing. This application was written in C# and architected by utilizing Weka 3.5.6 using the IKVM 0.36.0.5 bridge and using SPSS version 15.0.1 through the SPSS .NET API. Weka's J48 was selected to build and test classification models for each dataset.

Results

One-way ANOVAs were used to evaluate each of the five factors in Hypothesis I, namely the level of data quality for each of the four dimensions and for the level of problem complexity. Each was found to have significant main effects ($p < .01$, the criteria used for all statistical tests in this research). Confirmation of Hypothesis I was expected and a necessary step before proceeding.

Hypothesis II theorized the existence of significant differences in F-measures when two levels of data quality were compared while controlling for the effects of problem complexity. Thirty six one-way ANOVAs for each individual hypothesis in Hypothesis II showed significant main effects for all but one minor case. The direction of effects confirmed the significance of the component hypotheses in Hypothesis II; decreasing data quality levels lead to decreasing F-measures.

Subsequent analysis found a pattern of effects when data quality levels were varied. The same relative effect size and the same relative response in F-measures to changes in data quality were evident for each dimension. The smallest effect sizes were for varying levels of timeliness. Varying levels of completeness and accuracy had similar effect sizes, despite the difference in the algorithm used to generate them. Consistency stood out as the dimension in which varying levels produced the greatest effect sizes and the most rapid deterioration for decreasing levels. Figure 5 is a representative plot of mean F-measures as a function of data quality level for problems of medium complexity. The same pattern of data quality dimensions relative to each other could be seen in plots for other levels of complexity.

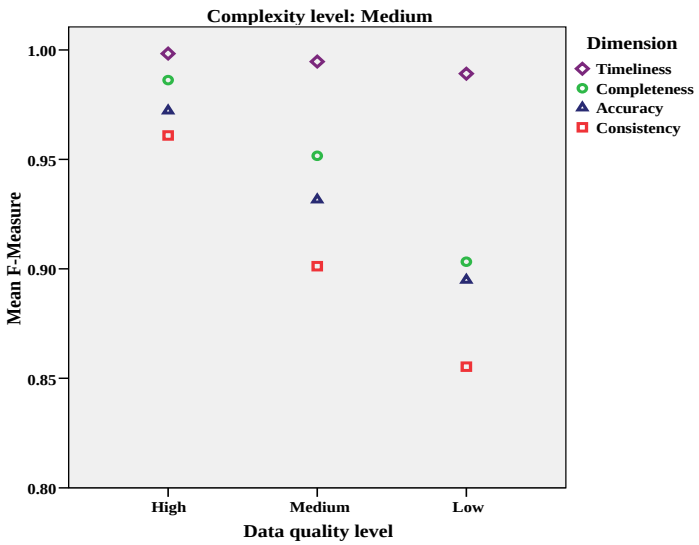

Figure 5: Comparison of data quality dimensions for problems of medium complexity

Hypothesis III theorized significant effects from comparisons of problem complexity levels while controlling for data quality. In evaluating its 36 individual hypotheses using one-way ANOVAs, problem complexity was found to have significant effects on data mining outcomes for all but two comparisons of accuracy, completeness, and timeliness. Decreasing F-measures in these three dimensions for decreasing levels of problem complexity confirmed Hypothesis III for the significant effects: increasing levels of problem complexity have a significant and negative effect on data mining outcomes.

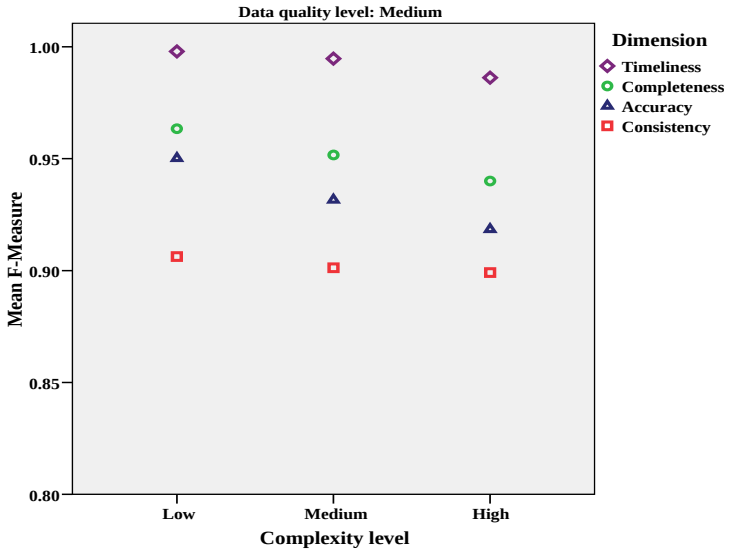

Figure 6: Comparison of data quality dimensions for medium data quality

An initially unintuitive result was that no significant effects were found for any comparisons of two levels of problem complexity when the dimension of consistency was held at one level. Reflecting on the results from Hypothesis II it became clear that not only did consistency have the largest effect size, but that F-measures degrade much more rapidly in response to decreases in consistency than any of the other three dimensions, a pattern evident in Figure 5 by the widening gap in F-measures between consistency and the other three dimensions as data quality decreases.

The outcomes of data mining are also more sensitive to changes in consistency than problem complexity. Correlations between the values of data quality and problem complexity metrics expressed as R^2 are shown in Table 1. Consistency explains more of the variability in F-measure than any other dimension. More importantly, even changes in consistency within the bounds of a single level of quality have a greater effect than changes between two levels of problem complexity as can be seen by the low R^2 of .005 for problem complexity when compared to consistency.

Dimension	Correlation (R^2) of F-measure with data quality	Correlation (R^2) of F-measure with problem complexity
Accuracy	0.689	0.126
Completeness	0.755	0.073
Consistency	0.884	0.005
Timeliness	0.131	0.233

Table 1: Coefficients of correlation between data quality and problem complexity by dimension

Because the method used to represent varying levels of consistency reduces the separation between classes it might be expected that this dimension is most related to problem complexity. It therefore might be expected that this dimension would have the least chance of having either significant main or interaction effects. This is not the case and the effects of small changes in consistency outweigh those of larger changes in problem complexity. This surprising result indicates that the lack of consistency, symptomatic of problems with referential integrity, has a much greater impact on data mining than does entropy or problem complexity.

Hypothesis IV hypothesized the existence of significant interactions between the levels of both data quality and problem complexity for each dimension. Two-way ANOVAs showed there are significant interaction effects between data quality and problem complexity for accuracy, completeness, and timeliness, but again not for consistency.

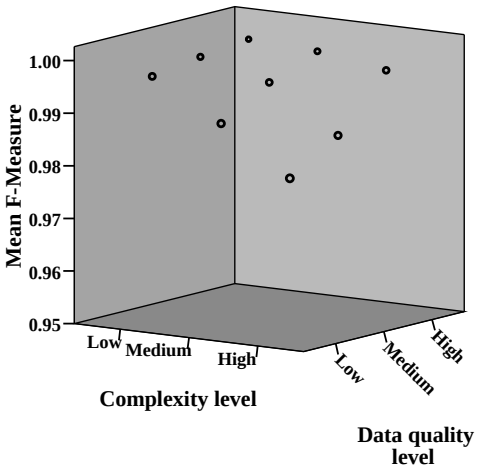

Figure 7: Plot of mean F-measures for varying levels of timeliness and complexity

The interaction effects for timeliness can be visualized in Figure 7. This plot shows mean F-measures decrease as either data quality decrease or problem complexity increase. It also shows that the combination of decreasing data quality and increasing problem complexity decrease F-measures more than either alone, as indicated by the non-linear drops in F-measure as either

factor degrades. Plots for the other two dimensions were similar; as with the main effects of data quality, the interaction effects with problem complexity were strongest for accuracy and completeness, and lowest for timeliness. Hypothesis IV was confirmed for these three dimensions by the direction of the interaction effects.

Again the results found for consistency stood apart from the other three dimensions, and again for the same reasons as found for consistency in evaluating Hypothesis III. The conclusions that can be made from the results found for consistency and from the findings of significant main and interaction effects are brought forward next.

Conclusions

For the first time, our research takes a comprehensive view of data quality and demonstrates that specific metrics for accuracy, completeness, consistency, and timeliness can be formulated and used to measure data quality and demonstrates that each of four dimensions data quality has a significant effect on data mining outcomes. Our research shows that decreasing levels of data quality have correspondingly significant and negative effects on those outcomes.

From our study it can be concluded that there is a pattern in the sizes of effects for all four dimensions. Timeliness has the smallest effects; next higher are accuracy and completeness with nearly equal effects, and consistency the largest. Whether considered at each level of complexity as data quality changes, or at each level of data quality when problem complexity changes, the ordering and relative size of effects from highest to lowest dimension stays the same.

For the first time we can conclude that consistency has larger effects on data mining outcomes than either of accuracy, completeness, or timeliness. The effects of consistency are strong enough to outweigh effects from varying levels of problem complexity both for main and interaction effects. This was an unexpected and important finding of our study.

March and Hevner [15] pointed to the areas of data quality and data mining as at the confluence of data warehousing and decision support systems and as particular challenges for research. Data warehouses are frequently used for data mining, and referential integrity problems in data warehouses have been found to be common [19]. Our metric for consistency is based on a representation of referential integrity and the results we found for this dimension are a step towards addressing those challenges emphasized by March and Hevner.

Our research confirms that there are significant interaction effects between each of three dimensions of data quality - accuracy, completeness, timeliness - and problem complexity. The findings of significant main and interaction effects along with finding of patterns in those effects have implications for building models of data quality as a manufactured product and models of trade-offs between data quality dimensions. These findings also have importance for creating representations of data quality through meta-data.

For the real world this study shows that a set of metrics to assess data quality can be used by practitioners to determine which dimensions of data quality will most likely influence their data mining outcomes, and to alert them to potential negative synergistic effects of data quality and problem complexity. These results can be used to choose from alternate sets of attributes or to prioritize data cleaning efforts in order to minimize those negative effects.

Our research can be extended in several ways. Interaction effects between the four data quality dimensions themselves warrant investigation. The datasets generated for this study had one class and one attribute variable; this schema could be expanded to include multiple attribute variables possibly with varying statistical distributions as a factor. Finally, although the J48 classification algorithm is widely used different classification algorithms could be evaluated. Each of these is a potentially productive area for new research.

References

[1] P. Atzeni, *Conceptual modeling-er 2004: Er 2004: 23rd international conference on conceptual modeling, shanghai, china, november 8-12, 2004: Proceedings*, Springer, 2004.

[2] D.P. Ballou and H.L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," *Management Science*, vol. 31, no. 2, 1985, pp. 150-162.

[3] D.P. Ballou and H.L. Pazer, "Modeling completeness versus consistency tradeoffs in information decision contexts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 1, 2003, pp. 240-243.

[4] R.A. Dillard, "Using data quality measures in decision-making algorithms," *IEEE Intelligent Systems and Their Applications*, vol. 7, no. 6, 1992, pp. 63-72.

[5] C. Fisher, E. Lauria, and C. Matheus, "In search of an accuracy metric " in *Proceedings of International Conference on Information Quality*, Cambridge, MA, 2007.

[6] M. Ge and M. Helfert, "A framework to assess decision quality using information quality dimensions," in *Proceedings of Proceedings of the 2006 International Conference on Information Quality*, Cambridge, MA, 2006.

[7] M. Ge and M. Helfert, "A review of information quality research," in *Proceedings of International Conference on Information Quality*, Cambridge, MA, 2007.

[8] P. Gomes, J. Farinha, and M.J. Trigueiros, "A data quality metamodel extension to cwm," *Proceedings of the fourth Asia-Pacific conference on conceptual modeling-*, vol. 67, 2007, pp. 17-26.

[9] J.M. Juran, *Juran on quality by design: The new steps for planning quality into goods and services*, Free Press, 1992.

[10] B.D. Klein, D.L. Goodhue, and G.B. Davis, "Can humans detect errors in data? Impact of base rates, incentives, and goals," *MIS Quarterly*, vol. 21, no. 2, 1997, pp. 169-194.

[11] R. Kohavi, N.J. Rothleder, and E. Simoudis, "Emerging trends in business analytics," *Communications of the ACM*, vol. 45, no. 8, 2002, pp. 45-48.

[12] K. Lakshminarayan, S.A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Applied Intelligence*, vol. 11, no. 3, 1999, pp. 259-275.

[13] W.Z. Liu, A.P. White, S.G. Thompson, and M.A. Bramer, "Techniques for dealing with missing values in classification," *Lecture Notes in Computer Science*, vol. 1280, 1997, pp. 527-536.

[14] J. Maddox, "Complicated measures of complexity," *Nature*, vol. 344, no. 6268, 1990, pp. 705.

[15] S.T. March and A.R. Hevner, "Integrated decision support systems: A data warehousing perspective," *Decision Support Systems*, vol. 43, no. 3, 2007, pp. 1031-1043.

[16] T.M. Mitchell, *Machine learning.* , Mac Graw Hill, 1997.

[17] A. Motro and I. Rakov, "Estimating the quality of databases," *Flexible query answering systems*, Vol. Volume 1495/1998, Springer, 1998.

[18] T. Oates and D. Jensen, " The effects of training set size on decision tree complexity," in *Proceedings of Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, Morgan Kaufmann Publishers Inc., pp. 254-262.

[19] C. Ordonez and J. García-García, "Referential integrity quality metrics," *Decision Support Systems*, vol. 44, no. 2, 2008, pp. 495-508.

[20] A. Parssian, "Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions," *Decision Support Systems*, vol. 42, no. 3, 2006, pp. 1494-1502.

[21] A. Parssian, S. Sarkar, and V.S. Jacob, "Assessing data quality for information products: Impact of selection, projection, and cartesian product," *Management Science*, vol. 50, no. 7, 2004, pp. 967.

[22] L.L. Pipino, Y.W. Lee, and R.Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, 2002, pp. 211-218.

[23] J.R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, 1986, pp. 81-106.
[24] T.C. Redman, "Data: An unfolding quality disaster," *DM Review*, vol. 6, 2004, pp.
[25] G. Shankaranarayanan and Y. Cai, "Supporting data quality management in decision-making," *Decision Support Systems*, vol. 42, no. 1, 2006, pp. 302-317.
[26] Y. Su and Z. Jin, "Assessment and improvement of data and information quality," in L. Al-Hakim, ed., *Information quality management: Theory and applications*, Idea Group Inc., 2007.
[27] J. Taylor and N. Raden, *Smart (enough) systems: How to deliver competitive advantage by automating hidden decisions*, Prentice Hall Professional Technical Reference, 2007.
[28] Y. Wand and R.Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, no. 11, 1996, pp. 86-95.
[29] R.Y. Wang, M.P. Reddy, and H.B. Kon, "Toward quality data: An attribute-based approach," *Decision Support Systems*, vol. 13, no. 3-4, 1995, pp. 349-372.
[30] R.Y. Wang, V.C. Storey, and C.P. Firth, "A framework for analysis of data quality research," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 7, no. 4, 1995, pp. 623-640.
[31] R.Y. Wang, M. Ziad, and Y.W. Lee, *Data quality*, Kluwer Academic Publishers, 2000.
[32] I.H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2005.
[33] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, "Enhancing data analysis with noise removal," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, 2006, pp. 304-319.