SUBJECTIVE EVALUATION OF PERCEPTION OF ACCURACY IN VISUALIZATION OF DATA

(Research-in-Progress)

Ahmed Abuhalimeh
IQ Program, University of Arkansas at Little Rock
aaabuhalime@ualr.edu

M. Eduard Tudoreanu
Information Science, University of Arkansas at Little Rock
metudoreanu@ualr.edu

Erich A. Peterson
Applied Science, University of Arkansas at Little Rock
eapeterson@ualr.edu

Abstract: This paper focuses on the human's perception of information quality and describes the results of a study on how accuracy is estimated for data shown through a visual representation. The subjective assessment of quality appears to be non-linear in relation to the actual degree of errors in the dataset. Users are sometimes unable to distinguish between datasets with different quality, and their ability to estimate is better for certain quality levels than for others. The study also shows that adding complementary information does not always help users to better assess the accuracy of the visualization, and thus of the data. The implication of these results is that, for subjective measures of quality, traditional statistical methods of assessing quality may need to be extended with additional methods to account for the non-linearity and the behavior of data integration.

Keywords: Data Quality, Information Quality, Subjective Quality

INTRODUCTION

In many applications the quality of information cannot be determined through algorithmic means and can only be assessed through subjective judgment. Some quality dimensions, such as believability, value-added or reputation, are intrinsically dependent on the human actor, while others may in certain situations become subjective. This may be the case for accuracy, a precise dimension when the exact value can be computed, and a subjective one when an exact value is not available or cannot be computed. Exact accuracy, for example, can be achieved when one checks the accuracy of some data against the balance of a bank account, and subjective accuracy is employed in the estimation of the amount of oil being spilled from a deep-water oil rig.

The term we introduce to describe quality measures that cannot be determined by a computer alone is subjective information quality or SIQ. Subjective information quality may not necessarily behave the same as the precisely computed measurements because they involve human factors and human psychology. Assessment of SIQ may be more application- and situation-specific, and rules for determining such quality may be different than statistical calculations. For example, people may find faults in data that is very accurate, and may find the combination of two poor data sources to be more than the sum or average of the parts.

The aspects of subjective information quality covered in this paper include how subjective rating varies with actual data quality, and how additional information supports better assessment. An orthogonal issue

is data integration. Decision makers, experts, and regular users often have to combine information from different sources to get a unified view of the information, or to help guide decisions on larger amounts of information from various sources. This process has become significant in a variety of situations both commercial and scientific. Combining data has become increasingly important as organizations strive to integrate an increasing quantity of internal and external information. Users must combine data for a variety of reasons. Some of those reasons are:

- to have more attributes;
- to get more detailed information for an attribute or item for different purposes and cases; or
- users cannot find an answer or a solution from a single dataset or data source.

Our results are based on a study on the perceived accuracy of weather data in the United States. The study employed data that can be easily judged by an average person living in the US. The data was obtained from the National Oceanic and Atmospheric Administration's (NOAA) website [1]. The investigators introduced a controlled amount of error in each visualization. Participants did not know the exact percentage of error introduced and were asked to estimate the visualization's overall accuracy.

The study also examined the effect of data integration by including visualizations with two panels, each conveying complementary weather data for winter and summer. The same amount of controlled error was thus presented to participants both as a single-panel and as double-panel visualizations. The additional panel itself had also the accuracy controlled and varied from completely accurate to fifty percent errors.

The study results showed that estimated accuracy is non-linear as a function of the actual accuracy, and that data integration may not always help users. Participants did not make constant estimation errors, nor did their estimation error increased or decreased with the actual accuracy. Multiple peaks and valleys are apparent, which suggests that people may not be able to distinguish between certain levels of accuracy, and that certain thresholds make accuracy estimation easier for a given number of actual error levels. With regard to data integration, the study found that introducing additional, error-free data, such as temperatures for summer in addition to those for winter, resulted in worse accuracy assessment than additional data with error. For this weather data set, we found the counterintuitive result that single datasets (for example winter only) are better estimated than datasets with double the amount of information (winter and summer).

The remainder of the paper is organized as follows: the next section discusses related work, followed by the experiment description and the results. The paper concludes with a discussion and future work.

RELATED WORK

Yang, Lee and Wang [3] present approaches that combine the subjective and objective assessments of data quality, however, their approaches do not ask for an estimation of the accuracy from participants. Their approaches are based on mathematical models, and focus on the data from one source. Our study provides visualizations techniques and aims to help in developing a method that will enable users to better estimate the quality of the data coming from different sources, since the statistical methods cannot determine the quality of the data in all situations, and especially for SIQ.

Motro and Rakov [4] provide a method for estimating the quality of data in databases. They propose to combine manual verification with statistical methods to arrive at useful estimates of the quality of databases. They propose a standard for rating information sources with respect to their quality that would help in our future work. Moreover, they show how to derive quality estimates for individual queries from such quality specifications. The authors focus only on two dimensions of data quality:

soundness/accuracy and completeness, and focus on the data extracted from one component. Our focus is on the human's perception of information quality in general, and describes how the accuracy is estimated for data shown through a visual representation. An important consideration is that the quality of information sources often varies considerably when specific areas within these sources are considered.

Lin and Hua [5] present a method for measuring data quality in data integration. They focus on believability and do not include the human interactions with quality, unlike our work.

Caballero, Verbo, Calero, and Piattini [7] proposed a Data Quality Measurement Information Model (DQMIM), which provides a standardization of the referred terms by following ISO/IEC 15939 as a basis. Their research deals with the concepts implied in the measurement process, and not with the measures themselves, they focus on two dimensions reliability and completeness. Our study focuses on the data quality dimensions and how people perceive those dimensions through visualizations which will help in developing a method that will help users to better estimate the data quality from different sources.

Peralta, Ruggia, Kedad, Bouzeghoub [8] project addresses the problem of data quality evaluation in data integration systems. They present a framework, which is a first attempt to formalize the evaluation of data quality. It is based on a graph model of the data integration system. In their project they only focus on data freshness and currency dimensions. Our study focuses on subjective dimensions such as accuracy.

Ballou, Chengalur-Smit, and Wang [9] research uses the relational algebra framework to develop estimates for the quality of query results based on the quality estimates of samples taken from the base tables. They do not discuss any data quality dimensions in their work; instead they measure the quality based on a specified condition, whether acceptable or not acceptable. The work assumes the quality of the data is not known. Our study assumes the data source is already asses and the quality is known for the subjective IQ dimensions and focuses on how people perceive subjective dimensions such as accuracy.

EXPERIMENT

Participants

The study was web-based, and was advertised to specific student groups in the information-related disciplines at the University of Arkansas at Little Rock and to colleagues of the authors. The study was open for about two weeks. 15 complete responses from 15 participants were identified. 3 responses were excluded because participants had selected the same answer for all questions. Participation was anonymous, and no information we stored could have been traced back to the participant. No incentives were offered.

Materials

Data

Data was obtained from NOAA [1] and included average temperatures for all US states broken down by season. Table 1 shows part of the dataset of the average temperature rates by state used in the study. Only winter and summer were included in our study.

STATE	WINTER	SPRING	SUMMER	FALL
ALABAMA	42.6	61.3	80.2	62.9
ALASKA	15.8	36.3	58.4	34.1
ARIZONA	51.7	60.8	86.5	70.5

ARKANSAS	40.1	61.4	82.4	63.3
CALIFORNIA	57.1	60.8	69.3	66.9

Table 1: Part of the weather data set employed in the study. All states were shown to participants

Equipment and software:
The computer hosting the web study was an Intel Xeon dual-core workstation running at 3.06 GHz with 3 GB of RAM and Windows 7 Professional (32-bit). The web pages were dynamically generated using the ASP .NET v4.0 framework, running on top of Internet Information Services 7.0. The web pages could be viewed on any network connection computer on campus, running the browser of the participant's choosing.

Methodology

The survey was broken down in six different pages, and participants could move to the next one by pressing a button. The first two pages were always presented in the same order, while the last four were presented in a random order determined in real-time by the web-server and our software every time a new browsing session (a new user) was establish.

The first page was a landing page with short instructions about the study. The second page, the warm-up, allowed the users to explore the features of the visualization type employed in the study. All visualizations in the study were created using the Many Eyes software [2] and consisted of a map(s) with the states within the US and an average temperature for each state. Many Eyes [2] is " an IBM research project and website whose stated goal is to enable data analysis by making it easy for laypeople to create, edit, share and discuss information visualizations" [7]. Some visualizations contained one map for one season (winter), and others two panels for two seasons (winter and summer).

The other four pages contained the actual visualizations that needed to be rated by the user. Table [2] shows the different pages presented to the participants, the maps shown on each, and the percentage each map was modified. For the Green Page the additional data was summer and was presented in the second panel, while for the Yellow and Blue Pages, the additional dataset was winter and shown first.

Page	Season(s) and % Modified
Red	Winter 6%
	Winter 12%
	Winter 25%
	Winter 50%
Green	Winter 6% and Summer 0%
	Winter 12% and Summer 0%
	Winter 25% and Summer 0%
	Winter 50% and Summer 0%
Yellow	Winter 25% and Summer 6%
	Winter 25% and Summer 12%
	Winter 25% and Summer 25%
	Winter 25% and Summer 50%

Blue	Winter 50% and Summer 6%
	Winter 50% and Summer 12%
	Winter 50% and Summer 25%
	Winter 50% and Summer 50%

Table 2: Pages shown to the user and the visualizations included in each. Note that the percentage shows the amount of error, which is the inverse of accuracy relative to 100%.

The datasets and the visualizations were generated before the study took place. A separate dataset was generated for each map/panel of the visualization by modifying the temperature in a certain percentage of the states for a given season. A special-purpose software was created for this task. The states were randomly chosen, and the temperatures were randomly generated within the minimum and maximum temperatures for that season that existed in the original data set. As such, all the modified temperatures still fall within some reasonable limits.

It is important to note that no dataset, except the 100% accurate (the original) set, was shown to the participants in more than one map/panel. Even for the same season and percentage of modification multiple datasets were generated with different states and different values.

After each visualization such as the one in Figure 1, participants were asked to assess the accuracy on a five level rating scale. For visualization composed of two panels (that is two maps), the instructions asked to rate the overall accuracy. The scale consisted in (1) Very Accurate (100% - 80%), (2) Accurate (80% - 60%), (3) Fairly Accurate (60% - 40%), (4) Inaccurate (40% - 20%), and (5) Very Inaccurate (20% - 0%).

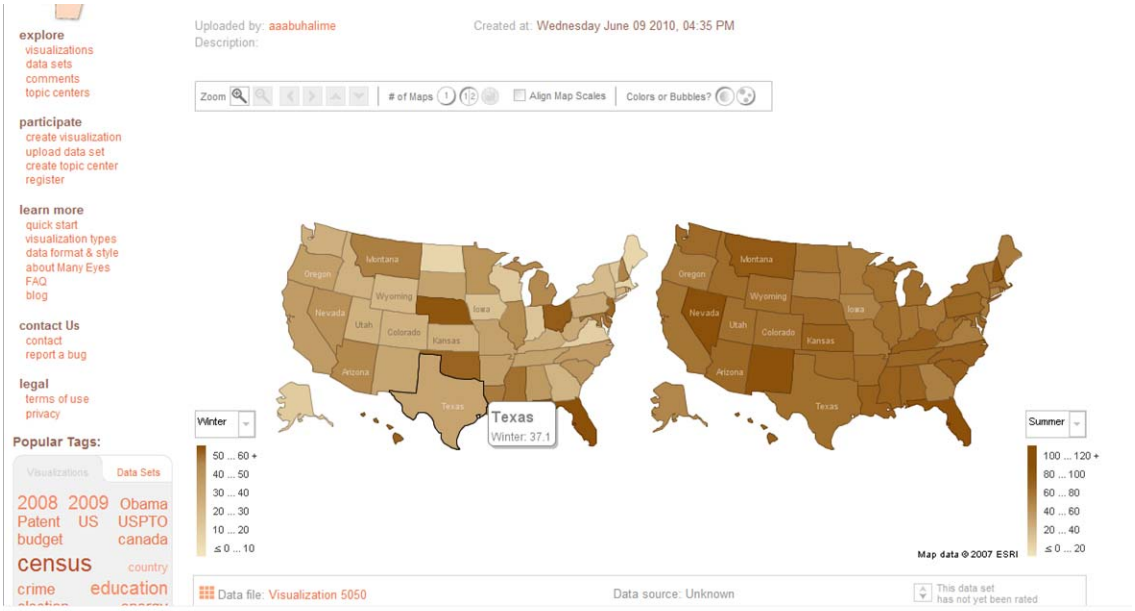

Figure 1: Snapshot of a webpage showing a visualization and question.

Hypotheses

The following hypotheses were considered:

A. Both the answers entered by the user and the amount of estimation error are dependent on the actual accuracy of the data shown in the visualization.
B. User answers and estimation error does not vary in a linear fashion with the actual accuracy of the data.
C. (C1) Adding additional information, such as a second season, changes both the answers of the participants and the amount of estimation error when compared to single-season data sets, and (C2) the more accurate the additional data is the more the overall subjective assessment of accuracy is improved.

Design

Two independent variables were considered: basic_accuracy, and additional_accuracy. The basic accuracy is one of the 94%, 88%, 75%, or 50%, and represents the quality of the data presented in at least one of the panels of each visualization. A webpage of the survey has four visualizations, one for each accuracy level.

Additional accuracy captures the quality of the data added in double panel visual representations. The additional accuracy is a constant within each webpage, but it varies from webpage to webpage. Possible values are -1 for single panel visualizations, and 100%, 75%, and 50% for double visualizations. For simplicity, the results will be reported using the either the accuracy of the single panel for simple visualizations (one of 94%, 88%, 75%, or 50%), or the average of the basic and additional accuracy for visualization composed of double panels (one of 97%, 94%, 87.5%, 84.5%, 81.5%, 75%, 72%, 69%, 62.5%, and 50%).

The dependent variable is user_estimation and it is one selection on a five level scale. The participants can choose one of the five, equal-size intervals that divide 0% through 100% accuracy. User's answer, as a measure, is independent of the actual accuracy, and the same answer for an accuracy of 94% can be significantly worse than for an actual accuracy of 50%. In order to quantify how exact participant's assessment was, we derived a metric from the user_estimation and average accuracy. The new metric, error is the difference between the average accuracy of a visualization and the closest edge of the interval answered by a participant. When the interval contains the average accuracy, the error is zero.

Results

The survey produced 15 complete answers, but the results only considered 12 because the other three appeared to resemble test submissions containing the same answer for all 16 questions. Overall, the analysis included 192 answered questions.

An ANOVA was performed for both the user_estimation and error. The average accuracy was found to be statistically significant factors: $F_{10,110}=2.01$, $p=0.0385$ for user estimation, and $F_{10,110}=2.26$, $p=0.0192$ for error. The same holds true for additional accuracy: $F_{3,33}=9.17$, $p=0.0001$ for user estimation, and $F_{3,33}=7.89$, $p=0.0004$ for error.

A Tukey pairwise analysis of the contribution of each additional type of visualization was performed. In the case of user estimation, significant differences were found between single-map (additional_accuracy = -1) and 100% accuracy additional panels (Adj. $p < 0.0001$). A weak statistical significance was found between single-panel and 75% (Adj. $p = 0.0692$). Another difference was found between 100% additional and 50% accurate additional maps (Adj. $p = 0.0440$). For error, single-panels (additional_accuracy = -1)

were significantly different than 100% additional views (Adj. p = 0.0002). Visualization containing 100% accurate additional panels were also found to be different than 50% and 75% ones, with Adj. p = 0.0080 and Adj. p = 0.0489, respectively.

The absolute values for user_estimation the error are depicted graphically. For simplicity, all graphs use the convention that the higher the bar the higher the error. Figure 2 shows the user answers and error recorded for various actual accuracy levels. Single visualizations were better estimated in term of error than double-panel ones (Figure 3). Figure 4 conveys the same in more detail and split by each level of accuracy for the additional panel. Figure 5 presents how both single and double-panel visualizations were assessed relative to the actual fault level in the additional data.

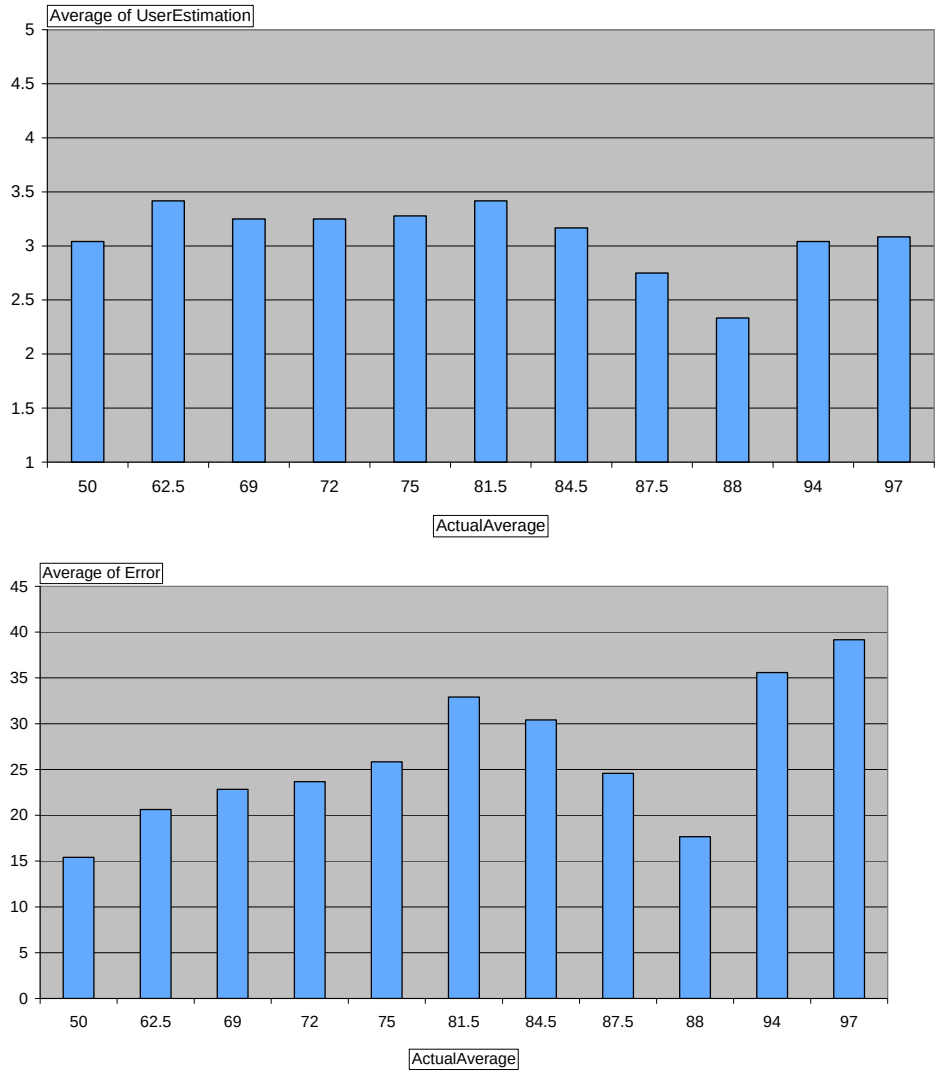

Figure 2: Average of user estimation (top) and error (bottom) per actual accuracy level. Note that, in the top panel, a value of 1 for user estimation means “Very Accurate (100%-80%)”, and 5 represents “Very Inaccurate (20% - 0%)”.

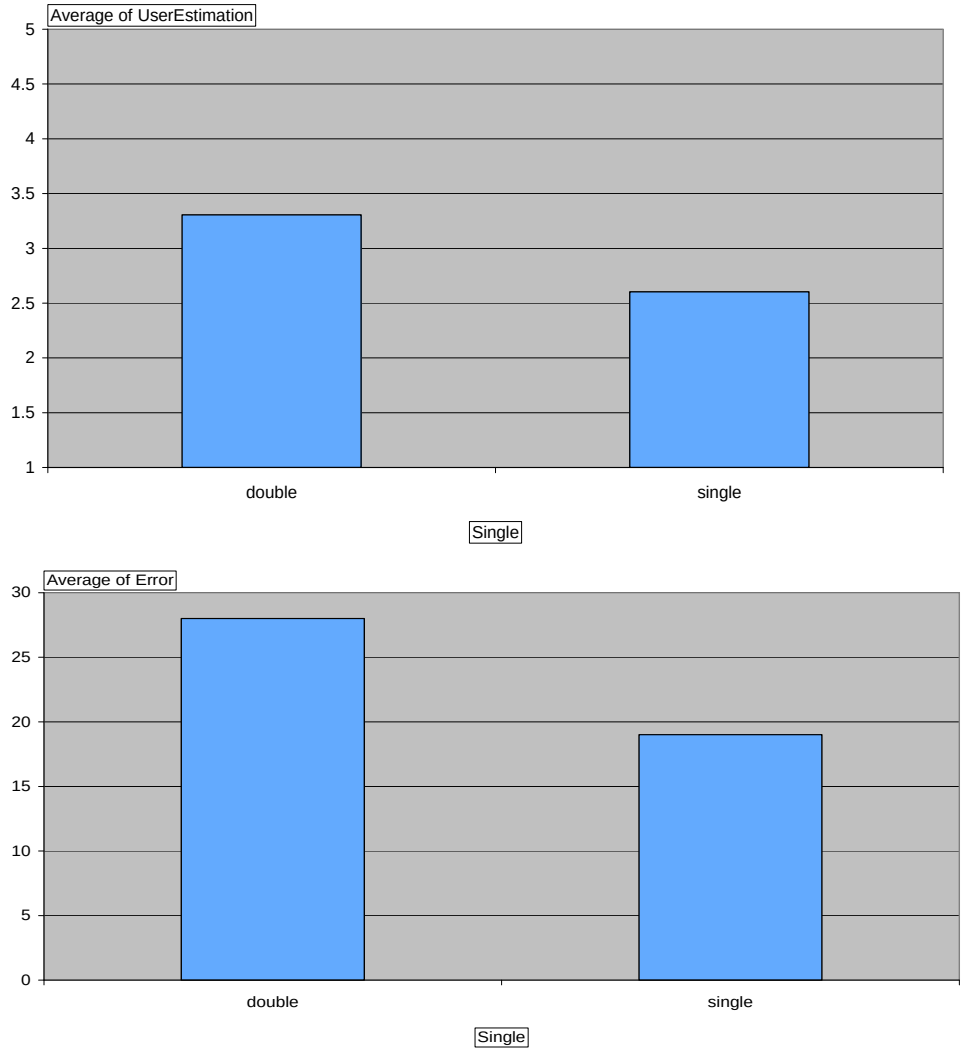

Figure 3: User Estimation (top) and error (bottom) for double- and single-map visualizations. Note in the top panel that "1" for user estimation means "Very Accurate (100%-80%)" and "5" represents "Very Inaccurate (20% - 0%)".

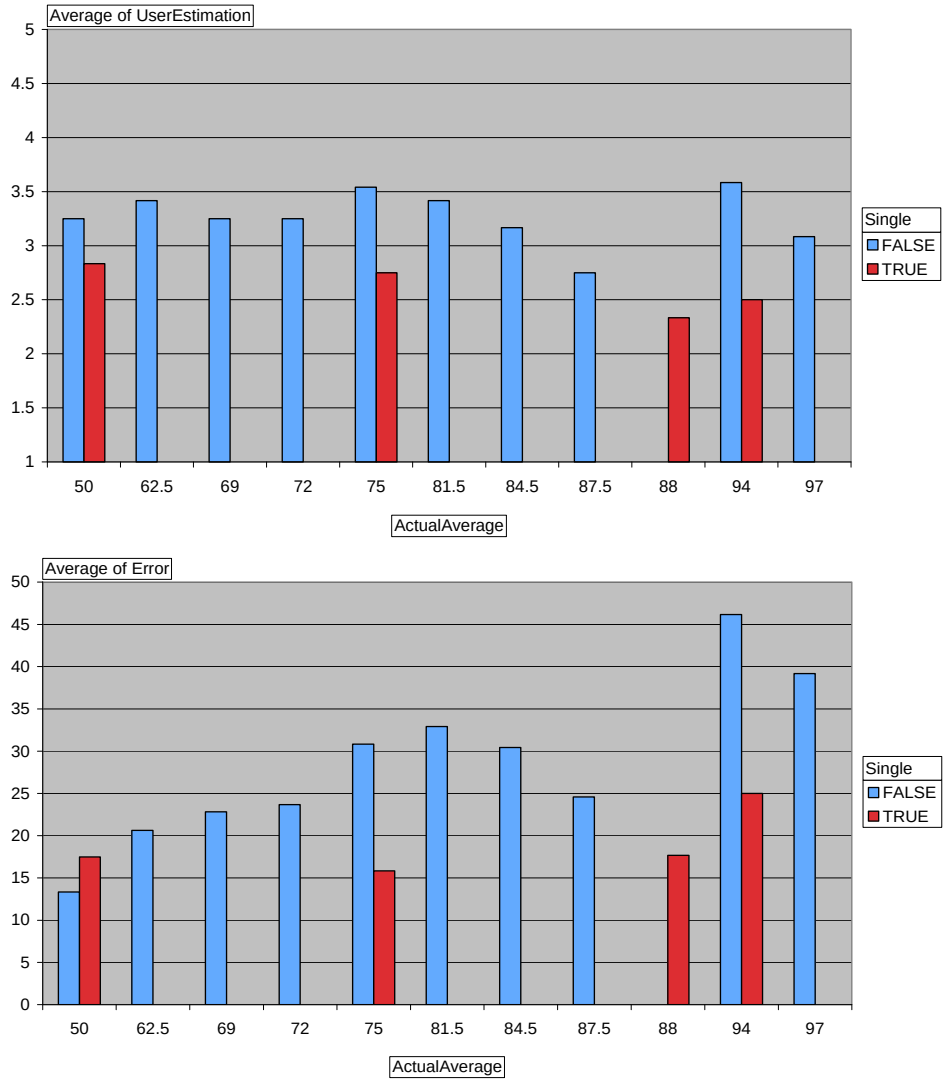

Figure 4: User Estimation (top) and error (bottom) for double- and single-map visualizations and divided per actual accuracy average. Note that 1 for user estimation means “Very Accurate (100%-80%)” and 5 represents “Very Inaccurate (20% - 0%).”

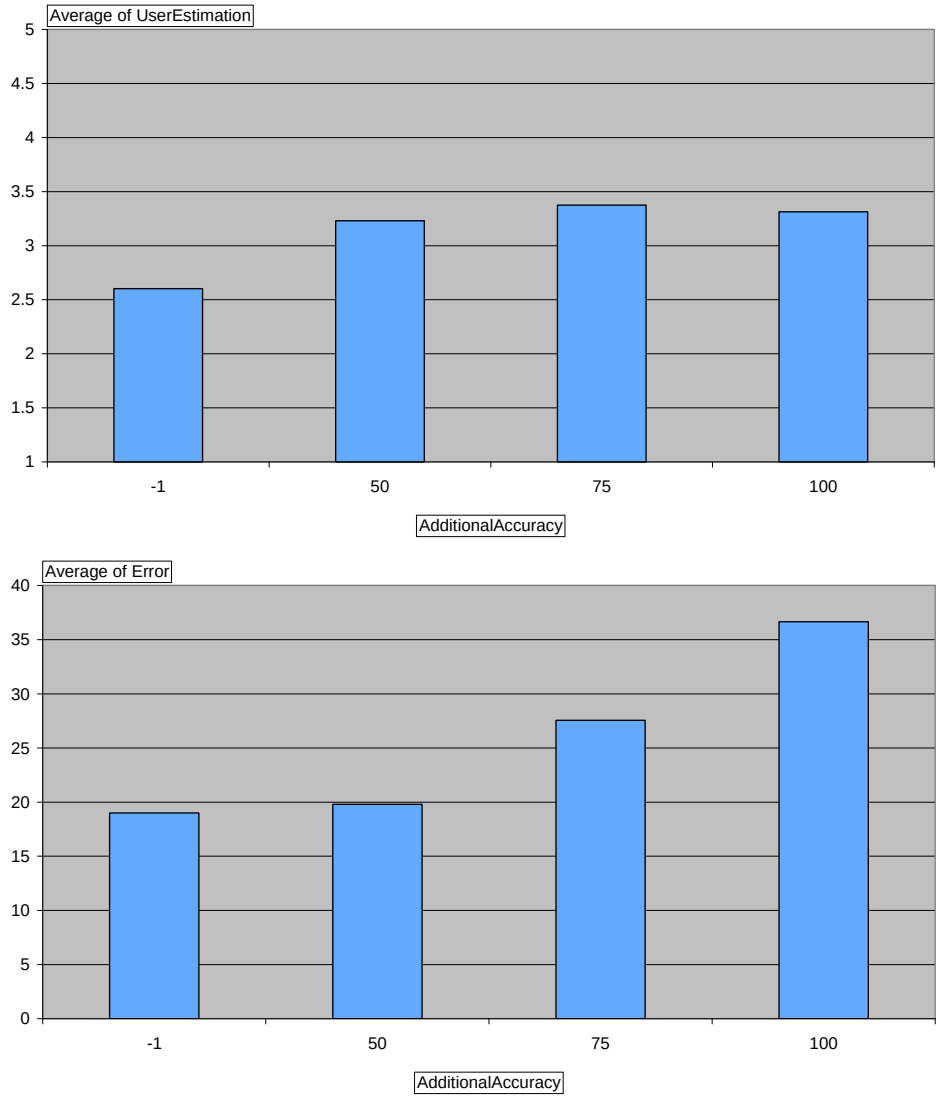

Figure 5: User Estimation (top) and error (bottom) as a function of the accuracy of the additional map (-1 denotes single visualizations). Note that 1 for user estimation means “Very Accurate (100%-80%)” and 5 represents “Very Inaccurate (20% - 0%)”.

DISCUSSION

Hypotheses A, B and C1 are confirmed by the experiment, but C2, the addition of accurate data helps subjective assessment, is not supported by the results. One explanation for the failure to observe C2 is that more information contributes to a task overload and participants performed worse. Initially, we thought that the addition of a winter or summer season would be able to provide more information about the minimum and maximum temperatures of each state. Then, if the one of the seasons was altered, the other season’s temperatures would help in identifying a too hot or cold temperature. Another way in which the additional season could help is by providing a visual pattern of temperature variations for a whole region of the US. Wide differences between the summer and winter pattern would be a warning that information in that region is not accurate. In the end, none of these suppositions were correct as shown by the failure to prove C2.

The results show that visualizations with single maps behave better than other types of visualizations (Figures 3-5), although no significant difference was found, on any of the dependent variables, between single-map and double maps in which one of the panels is 50% accurate. It appears that for this task users perform better with additional inaccurate information than with completely accurate one. This behavior would be difficult to incorporate in a statistical model for data integration. It may be that various thresholds may exist for when users are able to best use additional information.

There does not seem to be a linear dependence between the user assessment and the actual accuracy. The assessment is best around 88% and 50% (Figure 2), but it becomes worse around 100% and 80% (Figure 2). While "bad" estimation at 100% and "good" at 50% can be explained by its distance from the average of a completely random answer (that would be 50%), however there is no explanation for the behavior at 88% and 80%. Moreover, all of 80%, 88%, and 100% belong to a single answer interval: "Very accurate (100%-80%)". It is unclear why people think that almost accurate visualizations are worse than the ones that have 12% errors. It may be possible that people have thresholds of how they perceive visualizations, and that they also suspect that an error has been introduced even for very accurate visualizations.

FUTURE WORK

Assessing information quality is not an easy task and requires knowledge and awareness of the subjective and objective information quality metrics. Further studies may focus on additional levels of accuracy around the minimum and maximum values of Figure 2 to better discover any possible threshold. Such thresholds may also need to be determined for other tasks, data types, and data presentation methods.

Subjective assessment is not limited to accuracy, and our plans are to consider other SIQ dimensions and verify whether their behavior is similar to the subjective accuracy. Dimensions that are inherently subjective such as believability and value-added may lead to the development of a more complete theory of SIQ.

Any theory of SIQ may need to also consider the effect of data integration, an important topic in information and data quality. This study also showed that adding extra information is not always beneficial. Furthermore, for the cases when additional data is included, lower quality data may provide better support for subjective evaluation than higher quality data.

REFERENCES

[1] National Oceanic and Atmospheric Administration, www.nws.noaa.gov

[2] Many-Eyes Software. www.manyeyes.com

[3] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, Data Quality Assessment, Communications of the ACM, 2002, Volume 45 ,pp.211-118.

[4] Amihai Motro and Igor Rakov , Estimating the Quality of Data in Relational, Databases, Springer Berlin / Heidelberg, 1996, Volume 1495/1998, pp.298.

[5] MO Lin, ZHENG Hua , A Method for measuring data quality in Data Integration, International Seminar on Future Information Technology and Management Engineering, 2008 pp.525-527.

[6] Wikipedia, http://en.wikipedia.org/wiki/Many_Eyes.

[7] Ismael Caballero, Eugenio Verbo, Coral Calero, Mario Piattini, A DATA QUALITY MEASUREMENT INFORMATION MODEL.

[8] Verónika Peralta, Raúl Ruggia, Zoubida Kedad, Mokrane Bouzeghoub, A Framework for Data Quality Evaluation in a Data Integration System, 2004.

[9] Amihai Motro and Igor Rakov, Estimating the Quality of Data in Relational Databases, 1996.