Visual and Interactive Tools for Assessing Data Quality

Zachary H. Seeskin[1], Kiegan Rice[1]
[1]NORC at the University of Chicago

Abstract[*]
Federal statistical agencies are increasingly integrating data from multiple sources for analyses. These data sources often include data not collected for statistical purposes, including data used for public or private program administration. The Federal Conference on Statistical Methodology's 2020 report "A Framework for Data Quality" provides valuable guidance for standardizing data quality assessment practices. One area identified for further study is best practices for communicating quality, including graphical and interactive tools. We emphasize that exploratory data analysis is an important aspect of data quality assessment for understanding data sources' strengths and weaknesses. Applying methods informed by the data quality literature, we present graphical methods for exploring a data file and assessing the data quality components of accuracy, completeness of records, and comparability of the data over time and among subgroups. Further, we discuss the Data File Orientation Dashboard developed at NORC at the University of Chicago allowing users to interactively explore their data files, apply data quality analyses, and interpret the results to assess their data files.

Key Words: Exploratory data analysis; integrating data from multiple sources; administrative data; interactive data visualization; R

1. Introduction

The Federal Conference on Statistical Methodology (FCSM) recently released "A Framework for Data Quality" (FCSM 2020) providing strong guidance for federal statistical agencies and researchers on practices for evaluating data quality. This work expands on other recent data quality frameworks (Daas et al. 2011, Laitila et al. 2011, Iwig et al. 2013, Office for National Statistics UK 2013). The guidance applies to wide array of data sources, and in particular administrative data files used for the management of public and private programs as well as integration of administrative data files with other sources. These data sources are valuable for public policymaking and for statistical agencies often offering advantages regarding their size, support for granularity to study small subgroups, novel variables included, and strength for measurement. However, because the data are not collected for statistical purposes, a range of data quality concerns commonly arise, discussed in Seeskin et al. (2019). Thus, guidance for data quality assessment is critical so that users can assess the strengths and weaknesses of their data file to inform analytical decisions.

The FCSM Data Quality Framework includes an o*bjectivity* domain that is our focus in this article, which refers to "whether information is accurate, reliable, and presented in an accurate, clear and interpretable, and unbiased manner" (FCSM 2020, p. 3). When a user

[*] Please direct any correspondence about this article or questions about NORC at the University of Chicago's Data File Orientation Dashboard to Zachary Seeskin, Seeskin-Zachary@norc.org.

considers a new data file for statistical analysis, quantitative and visual methods for examining objectivity are particularly helpful to understand a file's suitability for planned analyses. The o*bjectivity* domain in the FCSM report includes *accuracy,* or the ability of a data product to support estimates close to their true values, *reliability*, or the consistency of results when the same phenomenon is measured in more than once under similar conditions, and *coherence*, which includes the ability to maintain consistency and comparability with other relevant data.

One recommended practice for statistical agencies and researchers from the framework is to "regularly identify threats to data quality for ongoing data collections, particularly when considering new sources of data for inclusion. Decisions on trade-offs among threats and mitigation measures should be considered in the context of the purpose of the data and all identified threats" (FCSM 2020, p. 5). That is, a proactive approach to assessing and ongoingly monitoring data quality is important for ensuring and maintain the objectivity of inferences coming from a data file.

Tools developed at NORC at the University of Chicago provide ready capabilities for structuring data quality assessments and identifying threats to data quality. We approach data quality assessment from the perspective of exploratory data analysis. Specifically, we view data quality assessment as a structured practice to explore a data file and identify potential data quality concerns for further investigation.

Two NORC tools focus on assessing the input data and provide guidance on interpretation for analysis decisions and potential threats to data quality, both incorporating methods from the data quality literature. The analyses included in the two tools are similar.

The Data File Orientation Toolkit is described in detail in Seeskin et al. (2019) and publicly available through a public GitHub repository (https://chapinhall.github.io/FSSDC/data-file-orientation-toolkit/). The toolkit is implemented using R Markdown, which readily enables sharing code and developing reports with result output. The scripts provide examples of data quality analyses with coding that the user can modify to the needs of their specific data sources. R Markdown is freely available to use with R and RStudio. For details about how to use the toolkit, please visit the GitHub repository.

The following sections describe the Data File Orientation Dashboard available internally at NORC, which allows for interactively analyzing and reviewing data quality analysis using R Shiny. First, we discuss how to use the dashboard before reviewing data quality methods incorporated in the dashboard.

2. Using the Data File Orientation Dashboard

The Data File Orientation Dashboard is a point-and-click interface allowing a user to run data quality analyses available without any programming. The first step is to load a data file to the dashboard on the upload page. Different formats are supported, and the user can select different kinds of delimiters for entries in the data file. The user will also see a preview of their file once loaded. Figure 1 shows an example of the results for a successfully uploaded file.

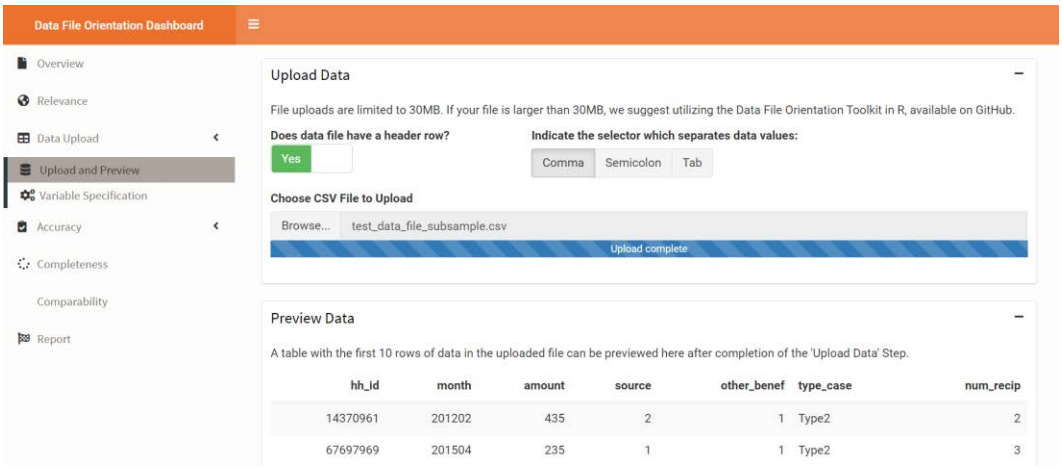

Figure 1: Data File Orientation Dashboard upload page.

Next, the user navigates to the variable specification page. The *data types* and *variable types* let the dashboard know which kinds of analyses to run based on the variables. *Data types* refer to the structure and format of the variable for analysis, including numeric, factor/or categorical, character, date, and logical variables. *Variable types* refer to the function of the variable in the user's planned analyses. These include variables indicating identification, time, key/dependent variables, domains including include categorical and continuous independent variables, and location/geographic variables.

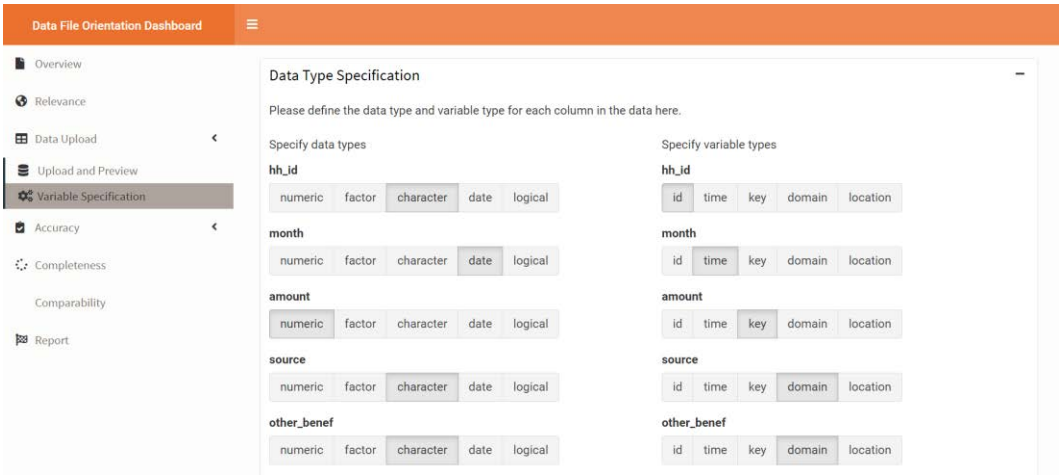

Figure 3: Data File Orientation Dashboard variable specification page.

Once the file is loaded and variable types are specified, the user can navigate through analyses oriented by data quality dimensions aligning with the *objectivity* domain of the FCSM data quality framework. The analyses are organized within data quality dimensions: *accuracy* of the data values, *completeness* of the data file, and *comparability* of values among groups or time periods of interest to compare. Each analysis includes written guidance for interpretation in the context of data quality assessment. The next section provides examples of analyses included.

3. Visual Tools to Assess Data Quality

Analyses in the dashboard provide different tabular and graphical output to examine variable distributions in the tool and assess the reasonable of data values. One analysis focuses on methods for outlier detection. Multiple visual methods are included to give the user a richer picture to review data quality. For example, the toolkit includes both boxplots and letter-value plots (Hofmann et al. 2017) for assessing data quality. Figure 3 shows an example analyzing a skewed variable of amount paid for building permits from the City of Chicago Data Portal. In this example, the variable is highly skewed. Without a variable transformation, the boxplot provides information about the distribution but does not provide a strong method for classifying outliers.

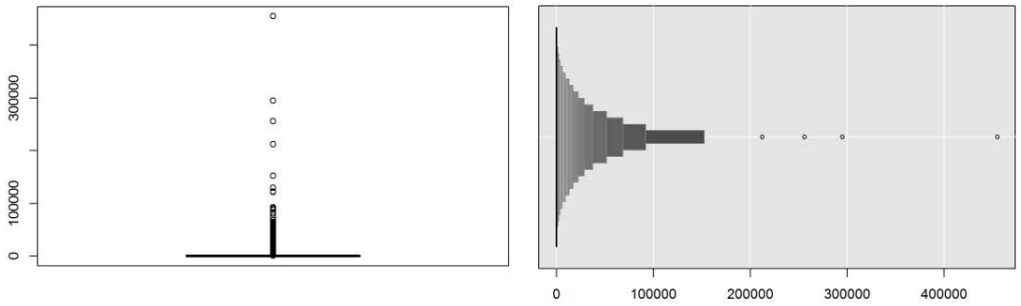

Figure 3: Boxplot (left) and letter-value plot (right) for a skewed variable.

The letter-value plot on the right provides a helpful depiction of the variable's distribution and identifies possible outliers for review. Letter-value plots, like boxplots, present the median with a vertical line and the interquartile range with the two boxes on either side represent the interquartile range. Thus, these boxes exclude the outer quartiles or 'fourths' of the distribution. The next two boxes exclude all but the lower and upper eighths, while the next two exclude all but the lower and upper sixteenths. The process iterates until certain stopping rules are achieved. Remaining points are plotted and may be reviewed as possible outliers.

The dashboard also runs analyses regarding *completeness* to assess both missing values for different variables and to assess possible groups of missing records from the data. As one example of the latter, Figure 4 shows a bar chart of the number of records by month in a simulated data file representing a cash benefits program. In this example, the number of records tends to decrease over time. Written guidance prompts the user for features to review from these visual analyses to understand the quality of their file, such as any gaps or jumps in the number of records over time and the reasonableness of the pattern in the context of the program.

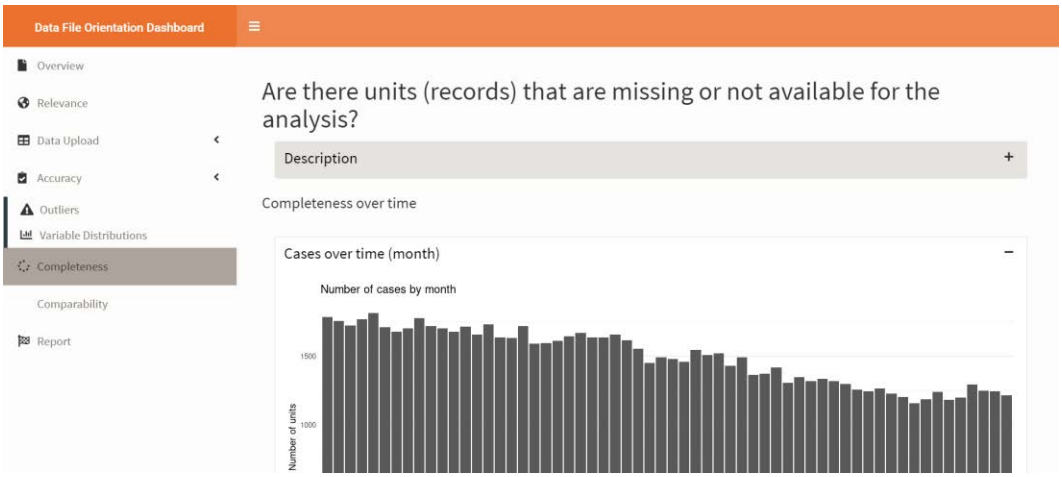

Figure 4: Bar chart of number of records by time period. Users may assess reasonableness of the pattern and potential for missing records.

Reviewing multivariate relationships can be enlightening to understand patterns in a user's data file. The toolkit includes tableplots (Tennekes et al. 2011), a data quality assessment method to examine these patterns and assess a data file's ability to support comparisons among different groups of a variable. Figure 5 shows a tableplot with five variables from the simulated data file for the state cash benefits program. All records are sorted by the values of a chosen variable in the table, in this example *type_case* the type of benefit, located in the first column. The sort variable is partitioned into percentiles with the *y* axis reflecting percentiles of the distribution of the sort variable. All other variables in the remaining columns have their values graphed at the different values of the sort variable in the first column. For continuous variables, the black line represents the mean value at the corresponding value range of the sort variable, and the shaded region represents values within one standard deviation of the mean. For categorical variables, different colors indicate the frequency of different categories at that value range of the sort variable. In this example, it is apparent that type of case has associations with *source*, *other_benef*, and *num_recip*. In addition, *amount* has different levels of variability depending on the category of *type_case*.

Figure 5: Tableplot to support examining multivariate relationship and comparability among groups.

Users may select different sort variables to review the multivariate patterns in the data file in different ways. Guidance is included as to features to review for investigating possible threats to data quality for the file and next steps for analyses.

After reviewing several visualizations described above, users have the option to download a report file containing summaries of their data file as well as the charts shown above. Written guidance for interpreting the visualizations and summary tables is included in the report file.

4. Conclusions

This article reviewed interactive and visual tools for assessing data quality. These tools have a critical role for supporting data quality assessment and can help disseminate best practices from data quality frameworks like those from FCSM. Well-designed tools can support users in guided data exploration of their files, enabling them to discover of features for further investigation regarding data quality and make informed analysis decisions.

Interactive tools such as the Data File Orientation Dashboard expand the user base for data quality analysis by removing the requirement of programming knowledge to assess data quality. The ability to perform a data quality analysis using a point-and-click interface can expedite the decision-making process regarding use of administrative data sources in research.

References

Daas, P., Ossen, S., Tennekes, M., Zhang, L.C., Hendriks, C., Haugen, K.F., Laitila, T., Wallgren, A, Wallgren, B., Bernardi, A. and Cerroni, F. 2011. List of quality groups and indicators identified for administrative data sources. First deliverable of WP4 of the BLUE-ETS project. http://www.pietdaas.nl/beta/pubs/pubs/BLUE-ETS_WP4_Del1.pdf

Federal Committee on Statistical Methodology. 2020. A Framework for Data Quality. FCSM 20-04. Federal Committee on Statistical Methodology.

Hofmann, H., Wickham, H. and Kafadar, K. 2017. Letter-value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, 26(3): 469-77. https://doi.org/10.1080/10618600.2017.1305277

Iwig, W., Berning, M., Marck, P. and Prell, M. 2013. Data quality assessment tool for administrative data. Prepared for a subcommittee of the Federal Committee on Statistical Methodology, Washington, DC. https://stats.bls.gov/osmr/datatool.pdf

Laitila, T., Wallgren, A. and Wallgren, B. 2011. Quality assessment of administrative data. Statistics Sweden. http://www.scb.se/statistik/_publikationer/ov9999_2011a01_br_x103br1102.pdf

Office for National Statistics UK. 2013. Guidelines for measuring statistical output quality. Office for National Statistics UK. https://www.statisticsauthority.gov.uk/wp-content/uploads/2017/01/Guidelines-for-Measuring-Statistical-Outputs-Quality.pdf

Seeskin, Z.H., Ugarte, G. and Datta, A.R. 2019. Constructing a toolkit to evaluate quality of state and local administrative data. *International Journal of Population Data Science*, *4*(1). https://ijpds.org/article/view/937

Tennekes, M., de Jonge, E. and Daas, P.J. 2011. Visual profiling of large statistical datasets. In New Techniques and Technologies for Statistics

conference, Brussels, Belgium. https://www.researchgate.net/profile/Edwin-De-Jonge/publication/228776940_Visual_Profiling_of_Large_Statistical_Datasets/links/5405add80cf2bba34c1d7510/Visual-Profiling-of-Large-Statistical-Datasets.pdf