Evaluating Commercial Data Quality

Commercial data can be evaluated based on its fitness for use and quality dimensions such as relevance, accessibility, interpretability, coherence, accuracy, and institutional environment (refer to **Data Quality Literacy Series 05: Understanding Administrative Data**). Common data quality issues that need attention include the following (for a more detailed explanation of each issue, see (Liu, 2020)):

- **Missing Values:** Missing values can occur due to skipped (or optional) questions in a questionnaire, data suppression for confidentiality, restrictions due to vendor agreements, data not being collected (e.g., not all retail stores collect random-weight data for fruits or vegetables), or other reasons. This can lead to confusion about the existence of the data or incomplete recording. A significant number of missing values can make a dataset unusable.
- **Data Errors** can happen in different forms including simple typos, arithmetic errors, coding errors, date errors, classification errors, etc. One way of identifying data errors is to compare different data sources.
- **Biases** are systematic errors caused by various reasons. Commercial datasets are prone to selection bias such as (1) **sampling bias**, where samples do not represent the population (e.g., opt-in market research panels); (2) **under-coverage bias**, where specific segments (e.g., senior citizens, low-income households, small or independent stores, or low-performing firms) are excluded or less represented in the sample); (3) **survivorship bias**, especially in financial datasets (e.g., Yahoo! Finance), when the dataset concentrates on collecting data about surviving stocks/firms while overlooking the data from delisted firms.
- **Inconsistencies** can happen from variable definitions to value format. For time-series data, consistency can be broken due to changes in questions or instructions in a questionnaire, classification system (e.g., North American Industry Classification System (NAICS)), geographic boundaries, eligibility, etc. Aggregate datasets from multiple private sources are susceptible to inconsistencies.
- **Discrepancies** between databases or datasets often arise from differences in coverage, definitions, coding policies, classifications, or errors. This can lead to the "database effect," where researchers may draw different conclusions based on the database they use.
- **Header Data** refers to data that only reflects the latest available value, which may not always be the most updated. Common examples of header data include company name, ticker symbol, stock exchange, industry code, and headquarters location. Header data in these data points can mislead time-series analyses by attributing data to incorrect or outdated identifiers, and distort cross-sectional studies by misclassifying entities based on incorrect groupings.
- **Standardization** improves data comparability across companies, time, and geography, but it can also result in understating or overstating the original outcome, leading to inaccuracies in certain prediction models.
- **Superseded Data** occurs when a dataset is revised or updated due to error corrections, restatements, or other changes. This can result in data downloaded at different times having different values. Substantial or systematic changes may raise concerns over data integrity.
- **Actual vs. Estimated Data:** Values in a dataset may be estimated based on models rather than exact numbers. This is common in financial data, especially for private companies' revenues or industry sizes in databases such as Dun & Bradstreet, Data Axle, or Bizminer. As a result, data sources often present widely divergent numbers in their estimates.
- **Reporting Time Issues** may occur when the dates that the data becomes available to the public are different from what researchers assume the date to be. Improperly recorded reporting time can lead to look-ahead bias or selection bias.
- **Misuse of Data** may occur when researchers improperly use the data as proxies or measurements, leading to unreliable research results. For example, data from databases that only cover public firms in an industry can be a poor proxy for calculating actual industry concentrations.

- **Lack of Transparency** is prevalent among database vendors. They often view their data collection methods or projection models as proprietary and are reluctant to disclose information about their data collection and management practices, as well as potential data problems and biases. To address this, probing vendors with quality-related questions is important in data acquisition.

To learn more, refer to **Data Quality Literacy Series**
07: Understanding Commercial Data
09: Commercial Data Quality: Conversation with the Vendors
10: Commercial Data Quality: Conversation with the Researchers

References

Liu, G. (2020). Data Quality Problems Troubling Business And Financial Researchers: A Literature Review And Synthetic Analysis. *Journal of Business & Finance Librarianship, 25*(3-4), 315-371.

Liu, G., Bordelon, B., Nagar, R., Sarti, J., Nguyen, U., & Boettcher, J. (2024). *Data Quality Literacy: A Guidebook.* Institute of Museum and Library Services (IMLS) Grant Project. https://doi.org/10.31219/osf.io/ruawm

The Knowledge Brief is compiled by Grace Liu, inspired by the National Forum presentation from Todd Hines and Patrick W. McLaughlin, reviewed by the IMLS Data Quality Literacy project team, and designed by Niko Galioto. This project was made possible in part by the Institute of Museum and Library Services [RE-252357-OLS-22].

Visit the project website to learn more!
https://www.dataqualityliteracy.org

INSTITUTE of Museum and Library SERVICES