

Modeling Statistical Comparisons in the Statistical Knowledge Network

Carol A. Hert
Research Professor
Syracuse University School of Information Studies
cahert@syr.edu

The Statistical Knowledge Network (SKN) is envisioned as a consortium of individuals and organizations supported by technologies that enable them to share and use statistical expertise and information (Marchionini et al., 2003). The SKN would provide integrated access to statistical information from a multitude of sources and enable use of that information through statistical literacy tools, appropriate interfaces, and so forth.

An important activity of users that must be supported in such a network is that of comparisons. These comparisons are of many types that include:

- Comparison across geographic units
- Comparison when there are definitional differences across concepts and variables
- Comparison across units of time (data collected/reported at different times, different aggregations (e.g., quarterly, monthly), and different reference periods)
- Comparison across different sources (websites, printed sources, etc.)
- Collection approaches (survey vs. census, household vs. establishment, etc.)
- Comparison across index values
- Terminology comparisons (same word-different concepts, same concept-different word)
- Deflated vs. real dollars
- Corrected vs. uncorrected data
- Preliminary vs. final release data
- Data with differing confidence intervals
- Seasonally adjusted data vs. non-adjusted data

The SKN should provide tools and explanations to enable non-expert users to understand what comparisons are appropriate, what aspects of a situation to consider when attempting a comparison, and perhaps suggest or implement courses of action.

A multi-stage approach is necessary to equip the SKN to facilitate the comparison process. First, the nature of the comparisons must be understood. Second, this knowledge of comparisons must be modeled in metadata and associated business rules for operating on the metadata. Finally, specific interfaces, statistical literacy tools, etc. need to be designed for the end-user. The first two components of this approach are reported in this poster.

In late 2003 and early 2004, the researcher conducted interviews with experts in one statistical agency, the United States Bureau of Labor Statistics. These experts had special expertise with comparisons (for example, several of them were employed to compare foreign labor statistics to those from the United States). A related set of website documents, research articles, and other documentation relating to comparisons was also analyzed. From these activities, the researcher developed a model of the comparison process and its dimensions. It was found that generally comparisons occur when a user wants to compare one's chosen "entity" (e.g., a cell in a table that may represent geographic, demographic, domain breakdowns) to:

- Others at one point in time,
- To itself over time,
- Or the combination of these.

While one wishes, in general, to compare actual statistics, this is not possible, since there is no way to determine whether those statistics are “true” due to both sampling and non-sampling error. An expert user resorts to several tactics:

1. Compare concepts and methodologies (including sample size, estimation procedures, coverage, universes, reference periods, other methods). In some instances, you might also need to have knowledge of the context of data collection and reporting efforts. If methodologies or concepts are found not to be comparable, users adjust, if possible, to achieve comparability.
2. Get a larger picture. An expert user may add additional measures/statistics or perform a variety of exploratory data analysis techniques to see how a given statistic “fits.” Rather than direct comparison, the task becomes finding the story that the variety of numbers tell. If a given number doesn’t “fit”, explanations are sought (e.g., checking for problems with data collection, seasonal adjustment (if appropriate), and/or understanding the domain/context in which the statistic sits).

The study also identified “rules of thumb” used by experts and particular “red flags” to be aware of when doing comparisons.

The next step of the work (spring 2004) is to identify the metadata elements associated with statistical information objects that can support comparison processes (see Denn, Haas, and Hert (2003) for information on the set of metadata elements under consideration). For example, some of those would be elements associated with methodological information such as sample, and data collection mode. Once these are identified and related to particular types of comparisons, business rules for manipulating those elements need to be determined. These would include rules for determining which metadata elements should be compared across object descriptions to identify objects appropriate for being compared, what the system should do if the comparison is not appropriate and so on.

The poster will provide a rich understanding of the user studies conducted to investigate the comparison process, the mapping of that statement to metadata elements and the associated business rules. The relationships to the larger vision of the Statistical Knowledge Network will also be considered.

Acknowledgements

This work was funded by the United States Bureau of Labor Statistics and National Science Foundation grant EIA0131824

References

- Denn, S.O, Haas, S.W., and Hert, C.A. (2003). Statistical metadata needs during integration tasks. Dublin Core 2003: Proceedings of the Annual Dublin Core Initiative Meeting. Pp. 81-90.
- Marchionini, G., Haas, S., Plaisant, C., Shneiderman, B., & Hert, C.A. (2003). Towards a statistical knowledge network. Proceedings of the National Conference on Digital Government Research 2003 (pp. 27-32).