REJOINDER

William E. Winkler, Energy Information Administration

Eli Marks' comments provide a valuable perspective to the overall objectives of matching procedures.

Just as the Fellegi-Sunter matching procedure contains computerized (automatic designation of matches and nonmatches) and manual (review of records designated for further manual followup) components, so does preprocessing contain computerized (minor reformatting, spelling standardization, string comparison) and manual (keypunch/transcription, major reformatting) components.

The respective roles of the two components are best exemplified by Newcombe et al. (1983, 1959, 1962). Newcombe's view is that computer procedures should be developed for the most routine and repetitive tasks. As knowledge of the characteristics of address files and coding techniques increases, computerized procedures can replace greater proportions -- possibly all -- manual components.

It is my experience that reasonably designed manual procedures are difficult and expensive to implement. This is because of high turnover rates and the necessity of training and constantly supervising personnel performing manual processing. Computerized procedures can have the benefit of being more cost-effective, consistent, and reproducible.

Both Marks and I note that the Census Bureau's ZIPSTAN software -- which is designed for files of individuals -- induced minor errors in files of businesses. In Winkler (1985), I show that ZIPSTAN's identification of address subfields can yield substantial improvements in the discriminating power of the Fellegi-Sunter matching procedure.

The cost in using ZIPSTAN was a few days of my time installing it. The alternative would have been to do nothing or develop manual procedures, set up computer files suitable for manual review, train individuals in computer login and manual review procedures, and have the individuals perform the review. Marks notes, if identifying individual subfields of the name and address involves "elaborate manual rearrangement and keying ..., substantial error is likely to be introduced, possibly as much as preprocessing removes."

I strongly agree that our understanding of "matching tolerances" needs to be improved. The purpose of my discussion of string comparators was to show the limitations of tolerances such as SOUNDEX, particularly SOUNDEX abbreviations of surnames used as sort keys during the blocking stage of matching. For files of businesses, I show (Winkler, 1985) that individual sort keys are generally not suitable for creating blocks containing most matched pairs. My solution is to apply independently multiple sort keys.

String comparison metrics, such as Jaro's string comparator, can only be efficiently used during the discrimination stage because they involve the comparison of corresponding strings from pairs of records. In my view, they offer the best opportunity for developing tolerances. How such tolerances fit in the framework of the Fellegi-Sunter model needs to be described and quantified.

REFERENCES

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," Science 130, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," Comput. Biol. Med. 13, 157-169.

Winkler, W. E. (1985), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, and Information Theory," Paper presented at the 1985 ASA Annual Meeting, Section on Survey Research Methods, August 4-8, 1985, Las Vegas, Nevada, (pp. 227-241 in this volume).