DISCUSSION

Eli S. Marks, Consultant

WINKLER

This paper discusses Bill Winkler's presentation on "Preprocessing of Lists and String Comparison."

Key factors in "Preprocessing of Lists" are:

1. The objectives of the system and the costs of various levels and types of matching error.
2. Costs of attaining a given matching accuracy level by preprocessing vs. other alternatives (e.g., suitably tailored "tolerances").
3. The nature of the matching system--manual, computerized, "mixed," etc.
4. How preprocessing is performed.

1. Objectives

The objectives of the system and the costs of matching error are intimately related. For example, if the objective is to estimate undercoverage of the U.S. census in each state, city, county, township, place, etc. for purposes of allocation of representation in Congress and state legislatures, city/county councils, etc. and for allocating federal and state funds to state and local jurisdictions, a uniform level of matching error everywhere is more important than the absolute level of matching error. Thus, preprocessing may have little value if its effect is to reduce the different types of matching errors by the same percentages in all jurisdictions. On the other hand, if preprocessing reduces urban matching error more than rural, it may be desirable or undesirable, depending upon whether the level of urban matching error without preprocessing is greater or less than the level of rural matching error without preprocessing.

2. Alternative Techniques

The objective of preprocessing (i.e., reduction of matching errors) can be attained by other means (e.g., the prescription of matching "tolerances"); and these techniques may cost less than preprocessing. For example, soundex coding is a form of "matching tolerance." That is, all disagreements of vowels and some disagreements of consonants are ignored in determining whether a pair of records match on the soundexed "identifier." One can, in fact, combine some preprocessing with tolerances (and, perhaps, other error-reducing techniques) to get a more efficient matching system than either can give alone. For example, one can prescribe standard abbreviations for the address suffixes "Avenue," "Street," "Road," "Drive," "Place," "Boulevard," etc., but also provide that an address match where the suffixes differ will be accepted unless there is another address match where the suffixes agree. For example, "Sutton Drive" would match "Sutton Road" unless either file contains both "Sutton Road" and "Sutton Drive."

Standard spelling of name and address may be achieved more accurately and more cheaply by controlling data collection, recording and "keying" (to put the data in machine readable form) than by preprocessing. This would, for example, avoid most of the errors of preprocessing by ZIPSTAN exhibited by the examples shown in the paper. Preprocessing errors can also be reduced or eliminated by other means, such as the clerical insertion of distinctive symbols to designate components of name and address, as outlined in Section 4 below.

It should be noted that selection of an "optimum matching strategy" is heavily dependent upon the type(s) of matching system(s) considered and that the choice of type of matching system is a vital part of the determination of "optimum matching strategy."

3. Kind of Matching System

The paper by Winkler notes that matching systems can be manual or computerized and implies that preprocessing is largely unnecessary for manual matching systems. I think his suggestion that individuals can usually determine accurately whether a pair of name and address records is actually a match or nonmatch is somewhat optimistic. Individuals can make this determination (so can a computer system), but how accurately depends on the kind of system. The great advantage of a competent human matcher operating in a properly designed matching system is the use of judgmental flexibility, provided, of course, he or she has good judgment and the matching rules permit him (her) to use that judgment (and I have seen many sets of matching instructions which do not). The great disadvantage of a well-designed manual matching system with competent matchers is the human matcher's slowness and the inevitable drop in efficiency in operating in a system which requires examining large masses of records; and not in lack of clear decision rules, inconsistency of application of decision rules, and nonreproducibility of results. All of the latter do occur, but can be adequately controlled in a well-designed matching system (although it is not easy!). However, humans cannot match the forte of the computer--its speed in examining large masses of data.

The solution to this problem is to let the computer do what it does well and let humans do what they do well. That is, design a mixed computer-human system, in which the computer handles the large mass of cases which can be classified as positive links or positive nonlinks, on a mechanical, routine basis. Carefully trained and well-motivated humans could then try to match the remaining cases,

using a "computer-interactive" system, where the human would specify a small class of possible matches and the computer would display the records in this class, until a positive link was found or there was adequate evidence that no such link existed.

4. Techniques of Preprocessing

Certain elements of preprocessing will unquestionably be valuable in any computerized matching system. In particular, it is important to develop some method so that the computer can quickly and accurately identify the various elements of the name and address: surname, house number, street name or number, first name, and the conventional prefixes and suffixes to name and address. If this involves elaborate manual rearrangement and keying of the name and address, substantial error is likely to be introduced, possibly as much as the preprocessing removes. The examples in the paper suggest that unaided computer formatting is also likely to introduce as much error as it removes. A solution may be something used in one of the earliest (1956) computerized matching systems, where clerks inserted a distinctive and computer-readable symbol in front of the components of name and address to be used in the matching; e.g., * before surname, # before house number, % before street name, $ before P. O. box number, @ before title, etc. After appropriate codes were placed in fixed fields, the symbols were deleted from the computer records.