A Largely Error-Free Register

Is Necessary for Survey Success

- Unmarked closures can be discovered, only if they are few. Specifically there can not be a lot of *not found* establishments.
- A few establishment discoveries, new or old, can be accommodated.
- Some errors in data can be fixed, but even in those cases some of the harm is already done.

Register Building and Updating

Applicable Country Situation

- No comprehensive list of industrial establishments
- Several lists that together provide a comprehensive list
- Small developing country
- CSO capability to maintain software
- System of Industrial Statistics is important
- First year start up technical assistance and extra resources provided.

- Lists are lists of establishments or enterprises, all the better if all of them are within scope. Some may be supplied by the CSO's regional bureaus.
- Register updating may be either continuous or cyclical
- There is much that is cumulative, *ie.* it does not need to be repeated each year.
- It is the start up that is hard.
- The procedure does not guarantee that missed establishments will not be too numerous, but can provide some checks.

Eight Steps

1. Getting useful lists from external and internal sources
2. Data cleaning and structuring to facilitate matching lists against the register and each other.
3. Reducing lists to establishments and enterprises within the scope of the register
4. Record linkage and matching

   Record linkage and matching is a process of finding establishments in the externally supplied lists and the CSO's internally preliminary list of new establishments that are *not* already in the register.
5. Identifying the candidates for addition to the register.
6. Prioritizing the candidates for field checking

7. Field checking

   Establishments are not added to register unless their existence and certain information about them is verified by direct contact.

8. Updating the register.

   Add the field checked candidates that turn out to be within scope.

4 Record Linkage - Blocking and Matching

Def 1. Two or more records *match* when they refer to the same establishment or enterprise.

Def 2. *Record linkage* is a procedure for linking records in one list with matching records in another list.

We should say *another* or *the same* list. It is quite useful in list processing to first match a list against it self in order to find and eliminate duplicate entries.

Def 3. *Blocking* is a procedure for linking a record in one list with *likely* matching records in another list.

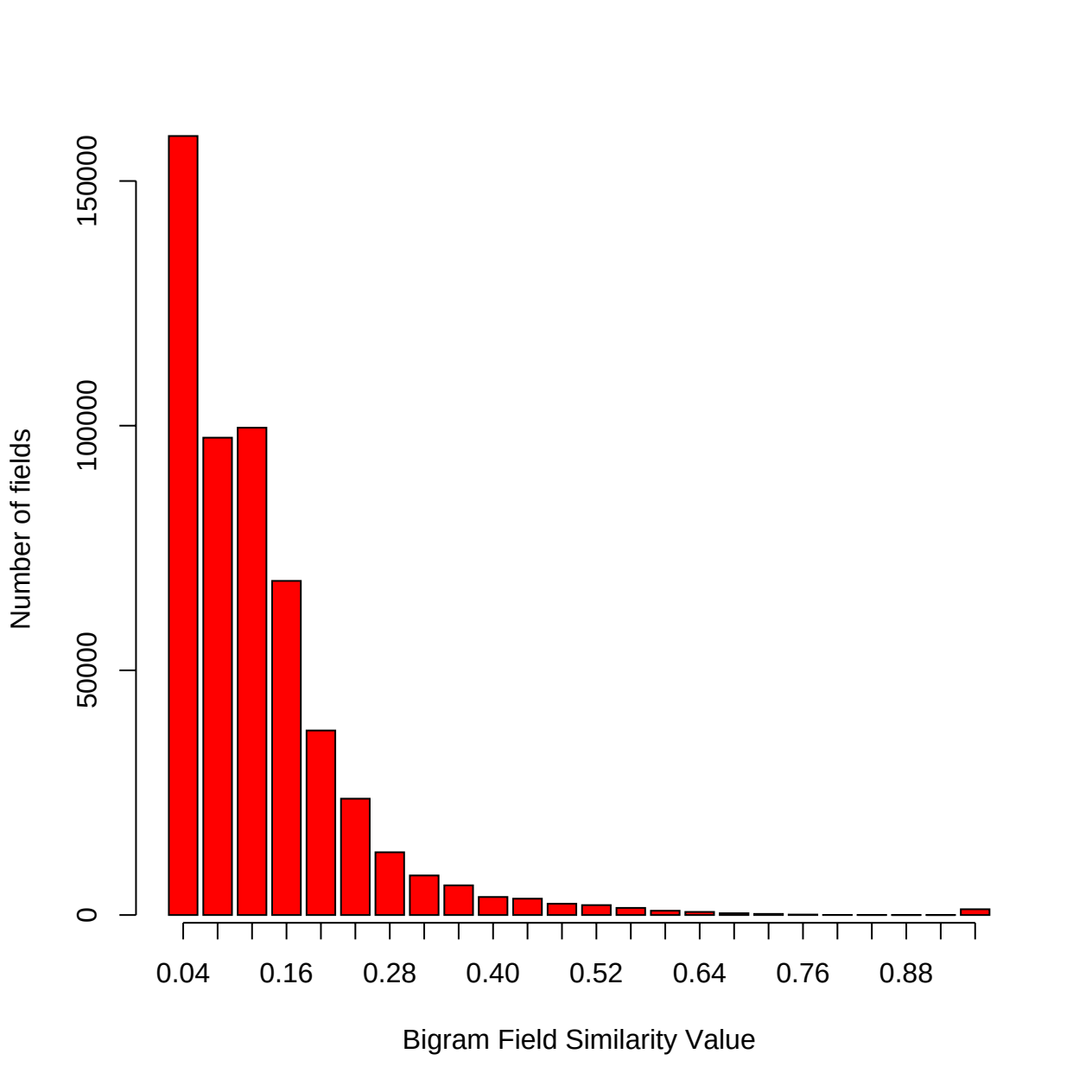
- Blocking is sort of the organizing idea of list processing.
- An algorithm compares a record s in list S with every record in list T and assigns, for each record in T, the likelihood that it is a match. We can write it as say, I_{sr}, where r runs over all of the records in T (and s over all S).
- Call the outcome of this computation the matching likelihood index, abbreviated *mli* .
- The *mli's* are not probabilities, but they are normalized to be nonnegative and have an upper bound of 1. It would be great if they were probabilities. Some matching systems do include probability estimates.
- The higher the *mli* the more likely a match.

More definitions

Def 4 The *blocking threshold* is an *mli* value that is fixed by the computer operator for the purpose of controlling the typical, or average, number of target records blocked.

Def 5 A record, t, from list T is *blocked* with respect to record s from the source list S when the *mli* for record s versus record t meets or exceeds a specific blocking value, namely the blocking threshold.

We can say s blocks t at an *mli* value of say v_{st}. It may be convenient to state *mli's* in terms of percentages. We could then say that record, s in list S blocks, and only blocks, the records $t_1, t_2, .., t_n$ in list T at an *mli* (a blocking threshold) of 85%.


Number of fields
0
50000
100000
150000
0.04
0.16
0.28
0.40
0.52
0.64
0.76
0.88
Bigram Field Similarity Value

Easy Matching

- If two lists are unique, there are no dups, so if s matches t, neither s nor t can match any other records.

This suggests a strategy.

- First match a list against itself to eliminate at as many duplicates as possible.
- Then in matching list S against T, do the easy matches first.

This will cause there to be fewer blocked records for for any given blocking threshold. Among other things this means starting with a high blocking threshold and then lowering it.

Skilful operators and supervisors will need a number of controls and reports to be effective.

Controls and Reports

1. The user can vary the matching threshold to vary the number of blocked records being presented.
2. The user can designate matches from among those presented.
3. The user can designate matches even when they are not presented.
4. The user can mark a pair of records as a non-match. This is a convenient way of reducing clutter when the blocking algorithm presents choices that the user knows are not matches.
5. The user can associate a comment with a match for future reference. This can be useful when the supervisor, who will do mainly quality control checking, needs to investigate a match that the clerk is not sure of.

6. The user can associate a comment with a designated non-match.
7. The user can undo match and non-match designations.
8. The user can mark records for delayed match processing. This is useful since as matching proceeds more and more potential matches will be taken off the table. Difficult cases will be made easier.

   In principle record matching is one-to-one. That is the lists are assumed to be unique with respect to establishments. Once a record in list S his matched to a record in list T, it cannot match any other record in list T.
9. The user can toggle some settings on and off. For example, bound records can released to be presented as potential matches, designated non-matches can be freed, records marked for delayed processing can be presented, and so forth.

10. The user can generate some progress and information reports showing: i) records in S that are matched versus not yet matched; ii) median blocks per remaining unmatched records in S for a user specified blocking threshold; iii) percentage of all records in S that are bound and likewise for T.

11. The user, unable to decide whether s matches t, may want see if t in fact makes a good match for some other record r in S. One approach is to delay the decision until all of the other records in S have been examined, but in some cases it may be more efficient get the software to switch the target and source and check all of the records in S against the specific record t in T. A good alternative match in S for t would take t off the table.

Field checking

- No establishment should enter the register without confirming its physical coordinates and whether it is within scope.
- In the first year the existing register will also need to be field checked.
- Subsequent updates for establishments missed in the annual survey, or not responding can be done on a rolling basis so that no establishment goes more than two years without completing a register questionnaire.
- Filling out register questionnaires using scripted phone calls may be possible.

New lists, updated list and continuity

- A good list processing/register updating system can swallow new data seamlessly *eg.* a new version of one of the lists may become available.
- Without going into detail, an updated list is compared with the old list. Changed and new records are identified. Those records that have not changed simply inherent whatever properties they have already acquired in the system. Changes in existing records may lead to a record that was not matched and not added to the register being either matched or, for whatever reason, being deemed a candidate (for addition to the register).
- Even a list containing a single establishment could be added to the system.

... continued

- Computationally, the hard work is calculating *mli's* . For each changed or new establishment in a list update or a new list, all of the *mli's* involving it need to be recalculated. For only a few cases this will not take long, for a couple of hundred or so it may need to be batched.
- Except for perhaps finding an infusion of new matches and some more good candidates for field checking the person at the console should notice nothing unusual.

Final Observations

- List processing at best can completely populate the industrial register. At worst it at least provides a partial solution. It may be, for example, that all the externally supplied lists can do is provide a check on the completeness of the CSO's own discovery effort. Those on the externally supplied lists that are not on the CSO's own list indicate, albeit perhaps imperfectly, how many and in what regions the CSO's is failing to discover establishments.
- If the internal and external lists are the same, then one or the other can be dropped.
- Very good software is essential to make the system work. A CS Pro module for matching and register management anyone?
- A well described set of written procedures is likewise essential.