12
Blocking

Suppose that for two files, A and B, of average size, the number of records in the product space $A \times B$ is too big for us to consider all possible record pairs. Because (1) only a small portion of the pairs in $A \times B$ are matching records and (2) there are 2^n possible comparison configurations involving n fields, drawing record pairs at random would require a sample size approaching all record pairs (for typical applications) to obtain sufficient information about the relatively rare matching pairs.

Blocking is a scheme that reduces the number of pairs of records that needs to be examined. In blocking, the two files are partitioned into mutually exclusive and exhaustive blocks designed to increase the proportion of matches observed while decreasing the number of pairs to compare. Comparisons are restricted to record pairs within each block. Consequently, blocking is generally implemented by partitioning the two files based on the values of one or more fields. For example, if both files were sorted by Zip Code, the pairs to be compared would only be drawn from those records whose Zip Codes agree. Record pairs disagreeing on Zip Code would not be on the same sub-file and hence would be automatically classified as non-matches (i.e., pairs in $\mathfrak{U}$).[1]

To illustrate this concept further, we consider an age variable. If there are 100 possible ages, then this variable would partition the database into 100 subsets. The first subset would be all of the infants less than one year of age. The second subset would be all those between ages one and two, etc. Such subsets are known as *blocks* or *pockets*.

Example 12.1: Calculating the number of non-matches

Suppose file A has 2,000 records and file B has 3,000 records. Further, suppose that there are no duplicate records within either file.

(1) How many possible record pairs, (a, b), are there in which $a \in A$ and $b \in B$?
(2) What is the maximum number of matches (a, b) where $a \in A$ and $b \in B$?

[1] In census studies, this grouping variable frequently identifies a small geographic area. This may well be the reason that this scheme is known as "blocking."

(3) If the number of matches attains its maximum value as computed in (2), what is the corresponding number of non-matches?

Solution

(1) The number of pairs in $A \times B$ is $2{,}000 \times 3{,}000 = 6{,}000{,}000$ – a large number indeed!
(2) The maximum number of matches is 2,000, the number of records in the smaller file.
(3) The number of non-matches is $6{,}000{,}000 - 2{,}000 = 5{,}998{,}000$. Thus, even the maximum number of matches is a relatively small proportion of the total number of pairs within $A \times B$.

Example 12.2: Calculating the number of pairs within an age group

Suppose that file A of Example 12.1 had 20 people at each of the ages from 0 to 99 and that file B of Example 12.1 had 30 people at each of the ages from 0 to 99. Suppose that a block consisted of all those people in either file A or File B at age 35. How many possible record pairs, $(a, b) \in A \times B$, are there in the block in which the recorded age on both records a and b is "35"?

Solution

There are 20 such individuals from file A and 30 from file B, so the answer is $20 \times 30 = 600$.

Hence, to do all pairwise comparisons within each of the one hundred blocks would require only $100 \times 600 = 60,000$ comparisons – only 1% of the 6,000,000 comparisons required to compare all of the records of file A to all of the records of file B without blocking.

12.1. Independence of Blocking Strategies

Blocking causes all records having the same values on the blocking variables to be compared. One consequence of this is that records which disagree on the blocking fields will be classified as non-matches. In particular, if the age field is used as one of the blocking variables and a record of file A has an erroneous age recorded, then that record would not be matched to the appropriate record of file B (unless the record on file B had a corresponding error). To circumvent this situation, we use multiple passes of the data.

Suppose that we ran a match in which the Soundex code of the last name was the sole blocking field on the first pass and the postal code of one's primary residence was the sole blocking variable on the second pass. If a record did not match on the first pass because the age was incorrect, then that record might still match on the second pass. Further passes could be made

until the analyst felt tha
of errors in the blockir
unlikely.

The blocking strategi
extent possible. For exa
and birthdate (year, mo
on last name, gender,
on first name and day
having errors in last na
would not be matched
second pass.

12.2. Blocking

To identify the matche
number of possible v
a low probability of
larger ones. It is pref
first pass. The record
the databases. Becaus
restrictive schemes ca
matches.

For example, gend
tions the file into tw
good blocking field t
usually more effectiv
it is time dependent
You want to use the
small segments so th
size.

Also, fields subject
fields. For example,
hence would not m
failure of two matc
would cause the rec
to be matched.

Blocking is a trad
pairs) and false no
matches because th
pass matching tech
minimize the effect

maximum value as computed in (2),
non-matches?

$\times 3{,}000 = 6{,}000{,}000$ – a large number

; 2,000, the number of records in the

$,000 - 2{,}000 = 5{,}998{,}000$. Thus, even
relatively small proportion of the total

pairs within an age group

20 people at each of the ages from 0
ıd 30 people at each of the ages from
of all those people in either file A or
:ord pairs, $(a, b) \in A \times B$, are there in
ɔth records a and b is "35"?

ı and 30 from file B, so the answer is

within each of the one hundred blocks
ɔmparisons – only 1% of the 6,000,000
e records of file A to all of the records

ıg Strategies

ame values on the blocking variables
; is that records which disagree on the
matches. In particular, if the age field
nd a record of file A has an erroneous
be matched to the appropriate record
a corresponding error). To circumvent
he data.
h the Soundex code of the last name
t pass and the postal code of one's
ıg variable on the second pass. If a
cause the age was incorrect, then that
pass. Further passes could be made

until the analyst felt that it was unlikely that matches would be missed because of errors in the blocking fields. Errors on three specified blocking fields are unlikely.

The blocking strategies for each pass should be independent to the maximum extent possible. For example, if a pair of files had last name, first name, gender, and birthdate (year, month, and day) fields, then the first pass could be blocked on last name, gender, and year of birth. The second pass could be blocked on first name and day and month of birth. Consequently, matching records having errors in last name (e.g., a maiden and a married surname), for example, would not be matched in the first pass, but would probably be matched in the second pass.

12.2. Blocking Variables

To identify the matches efficiently, blocking variables should (1) contain a large number of possible values that are fairly uniformly distributed and (2) have a low probability of reporting error. Smaller blocks are more effective than larger ones. It is preferable to use highly restrictive blocking schemes in the first pass. The records that match in the first pass can then be removed from the databases. Because most records that match will match on the first pass, less restrictive schemes can then be used in subsequent passes to identify additional matches.

For example, gender, by itself, is a poor blocking field because it only partitions the file into two sub-files. Even a field such as age is not a particularly good blocking field because (1) it is usually not uniformly distributed, (2) it is usually more effective to partition large files into more than 100 subsets, and (3) it is time dependent so it often disagrees on pairs of records that should match. You want to use the blocking to partition the database into a large number of small segments so that the number of pairs being compared is of a reasonable size.

Also, fields subject to a high probability of error should not be used as blocking fields. For example, apartment number is widely misreported or omitted and hence would not make a good blocking field. Accuracy is crucial because a failure of two matching records to agree on the values of the blocking field would cause the records to be placed in different blocks and thus have no chance to be matched.

Blocking is a trade-off between computation cost (examining too many record pairs) and false non-match rates (classifying matching record pairs as non-matches because the records are not members of the same block). Multiple-pass matching techniques, using independent blocking fields for each run, can minimize the effect of errors in a set of blocking fields.

12.3. Using Blocking Strategies to Identify Duplicate List Entries

Winkler [1984] describes a study that evaluated methodologies for accurately matching pairs of records – within a single list of records – that are not easily matched using elementary comparisons.

The empirical database he used was a part of a list frame that the Energy Information Administration (EIA) of the US Department of Energy employed to conduct sample surveys of sellers of petroleum products. The list was constructed from (1) 11 lists maintained by the EIA and (2) 47 State and industry lists containing roughly 176,000 records. Easily identified duplicates having essentially similar name and address fields were deleted, reducing the merged file to approximately 66,000 records. This was more suitable for Winkler's study because he was concerned with accurately identifying 3,050 other duplicate records that only had somewhat similar names and addresses, and so were more difficult to identify.

Winkler [1984] used the following five blocking criteria in his study:

1. The first three digits of the Zip Code field and the first four characters of the NAME field in its original form – that is, not parsed.
2. The first five digits of the Zip Code field and the first six characters of the STREET name field.
3. The 10-digit TELEPHONE number.
4. The first three digits of the Zip Code field and the first four characters of the longest substring in the NAME field.
5. The first 10 characters of the NAME field.

Table 12.1 summarizes the results of applying these five criteria to the list of interest. These criteria were the best of several hundred considered.

As we can see from Table 12.1, the best blocking criterion (criterion 4) failed to identify 702 or 23.0% of the 3,050 duplicates on the list. The reason criterion 4 works best is that the NAME field does not have subfields (generally words)

TABLE 12.1. Number of matches, erroneous matches, and erroneous non-matches using a single blocking criterion

Blocking criterion	Matched with correct parent (match)	Matched with wrong parent (erroneous match)	Not matched (erroneous non-match)	Percentage of erroneous non-matches
1	1,460	727	1,387	45.5%
2	1,894	401	1,073	35.2
3	1,952	186	1,057	34.7
4	2,261	555	702	23.0
5	763	4,534	1,902	62.4

to Identify Duplicate

ıated methodologies for accurately
list of records – that are not easily

art of a list frame that the Energy
Department of Energy employed to
n products. The list was constructed
nd (2) 47 State and industry lists
identified duplicates having essen-
deleted, reducing the merged file
more suitable for Winkler's study
identifying 3,050 other duplicate
es and addresses, and so were more

ocking criteria in his study:

1 and the first four characters of the
, not parsed.
1 and the first six characters of the

1 and the first four characters of the

d.

ing these five criteria to the list of
ral hundred considered.
locking criterion (criterion 4) failed
ates on the list. The reason criterion
ot have subfields (generally words)

hes, and erroneous non-matches using a

Not matched (erroneous non-match)	Percentage of erroneous non-matches
1,387	45.5%
1,073	35.2
1,057	34.7
702	23.0
1,902	62.4

TABLE 12.2. Number of matches, erroneous matches, and erroneous non-matches using combinations of blocking criteria

Blocking criteria	Matched with correct parent (match)	Matched with wrong parent (erroneous match)	Not matched (erroneous non-match)	Percentage of erroneous non-matches
1	1,460	727	1,387	45.5%
1–2	2,495	1,109	460	15.1
1–3	2,908	1,233	112	3.7
1–4	2,991	1,494	39	1.3
1–5	3,007	5,857	22	0.7

that are in a fixed order or in fixed locations. Consequently criterion 4 is able to match fields of records having the following forms:

John K Smith
Smith J K Co

Criterion 3 (TELEPHONE) only produced 186 erroneous matches and only missed 1,057 actual matches. Criterion 5 (first 10 characters of NAME) had both the highest number of erroneous matches, 4,534, and the highest number of erroneous non-matches, 1,902.

Winkler [1984b] next considered combinations of these blocking criteria. His results are summarized in Table 12.2.

Criteria 1 and 2 are applicable to all EIA lists because all such lists have identified NAME and ADDRESS fields. As many lists obtained from sources external to EIA do not have telephone numbers, criterion 3 is not applicable to their records. As a number of EIA lists have consistently formatted NAME fields, criterion 4 will yield little, if any, incremental reductions in the number of erroneous non-matches during the blocking process. Finally, criterion 5 should not be used as a blocking criterion. Although the inclusion of criterion 5 reduces the number of erroneous non-matches from 39 to 22 and increases the number of correct matches by 16, it increases the number of erroneous matches by a staggering figure of 4,363 (= 5,857–1,494).

Results from using criteria 1 and 2 of Winkler's study are summarized in Table 12.3.

TABLE 12.3. Number of duplicate pairs identified using criteria 1 and 2

Criterion 1	Criterion 2		Total
	Present	Absent	
Present	859	601	1,460
Absent	1,035	?	
Total	1894		

Assuming that the necessary assumptions hold, we use the basic Lincoln–Peterson capture-recapture estimator of Section 6.3.1 to estimate the number of records in the blank cell and thereby estimate the total number of pairs of duplicates on the list of records:

$$\hat{N}_{LP} = \frac{x_{1+}x_{+1}}{x_{11}} = \frac{1460 \cdot 1894}{859} = 3{,}219.$$

This is reasonably close to the actual value of 3,050 duplicate records in the list.

12.4. Using Blocking Strategies to Match Records Between Two Sample Surveys

In the next example, we use blocking strategies to efficiently match records between two record files. The larger file is the main file from the 2000 Decennial Census in the United States. This file has records on approximately 300 million individuals. The other file is the main Accuracy and Coverage Evaluation (ACE) file of approximately 750,000 individuals. The ACE is the 2000 version of the 1990 Post-Enumeration Survey (PES) but is twice as large as the 1990 PES. The 2000 Decennial Census file was constructed using optical scanning equipment to convert hand-written information to electronic form. The ACE data were collected via a Computer-Assisted Personal Interviewing (CAPI) interview. The ACE file is matched against the Census file by census blocks[2] in order to determine the extent of the overlap of the two files. The overlap was then used to estimate the undercount and over-count of the 2000 Decennial Census. The estimates that the Census Bureau produces by census blocks are ultimately used for reapportionment of Congressional districts and for Federal revenue-sharing.

Example 12.3: Using the 2000 ACE to evaluate alternative blocking strategies

After an exhaustive effort, the Census Bureau was able to determine that 606,411 records in the ACE file matched records within the 2000 Decennial Census file. This provided the Census Bureau (see Winkler [2004]) an opportunity to determine the efficacy of various blocking strategies. The results of using each of 11 blocking criteria, separately, are summarized in Table 12.4.

Except for Criterion 7, each criterion individually identified at least 70% of the matched pairs. Criterion 7 was designed to identify the matched pairs in which the first names and last names are switched. This accounted for 5.2% of the matched pairs.

The 11 blocking criteria listed above identified all but 1,350 matched pairs. The four best criteria – 1, 3, 11, and 9 – identified all but 2,766 matched pairs; while the five best criteria – 1, 3, 11, 9, and 8 – identified all but 1,966.

[2] Each census block consists of approximately 70 households.

TABLE 12.4. Number of matches identified by each blocking criterion

	Blocking criterion	Number of matches	Proportion of matches (in %)
1	Zip Code, First Character of Surname	546,648	90.1
2	First character of surname, First character of first name, Date of birth	424,972	70.1
3	10-digit telephone number	461,491	76.1
4	First three characters of surname, Area code of telephone number, House number	436,212	71.9
5	First three characters of first name, First three numbers of Zip Code, House Number	485,917	80.1
6	First three characters of surname, First three numbers of Zip Code, Area code of telephone number	471,691	77.8
7	First character of surname = First character of first name (2-way switch), First three digits of Zip Code, Area code of telephone number	31,691	5.2
8	First Three Digits of Zip Code, Day of Birth, Month of Birth	434,518	71.7
9	Zip Code, House Number	514,572	84.9
10	First three characters of surname, First three characters of first name, Month of Birth	448,073	73.9
11	First three characters of surname, First three characters of first name	522,584	86.2

Some of the matches that were not identified involved children living in a household headed by a single or separated mother. The children were listed under different last names in the two files. Their date of birth was missing in the ACE file and the street address of their residence was missing in the Census file. The records on these children also had names with a high rate of typographical error that may be at least partially due to scanning error.[3] Table 12.5 illustrates the nature of these typographical problems.

TABLE 12.5. Examples of typographical problems that cause matches to be missed (artificial data)

Relationship to head of household	Record in census file		Record in ACE file	
	First name	Last name	First name	Last name
Head of Household	Julia	Smoth	Julia	Smith
Child #1	Jerome	Jones	Gerone	Smith
Child #2	Shyline	Jones	Shayleene	Smith
Child #3	Chrstal	Jones	Magret	Smith

[3] In 2000, names were captured (i.e., entered into computer files) via a procedure that scanned hand-written text and converted it to characters. Prior to 2000, names were not captured in Decennial Censuses.

The names on the childrens' records that should be matched have no 3-grams in common. Here, we say two names have a 3-gram in common if any three consecutive letters from the name on one record appear in the corresponding name on the other record. Consequently, it is unlikely that such matched pairs could be linked through any computerized procedure that uses only the information in the Census and ACE files.

12.5. Estimating the Number of Matches Missed

If an estimate of the number of matches missed when comparing two lists is required, then the lists can be sampled directly. However, even if very large samples are selected, Deming and Gleser [1959] show that the estimated variances of the error rate (i.e., the proportion, p, of missed matches) often exceed the values of the estimates. This is particularly true when the error rate, p, is small.

If samples are not used, then capture–recapture techniques can be used to estimate the number of matches missed, as illustrated in the example of Section 12.3. Winkler [2004; Section 4] uses sophisticated (capture–recapture) log-linear models to estimate the number of matches missed in the problem considered in Example 12.3. Based on the capture–recapture estimates, Winkler estimated that no more than 1 in 10^{11} pairs of the remaining 10^{17} (300 million $\times$ 300 million) pairs that were not in the entire set of pairs obtained via blocking would be a match.

12.6. Where Are We Now?

In this chapter, we showed how blocking strategies could be used in record linkages to limit the number of pairs that need to be compared while at the same time minimizing the number of matches that are missed. In the next chapter, we discuss string comparator metrics. These can be used to account for minor data-entry errors and thereby facilitate record linkages. The topic of string comparator metrics is the last enhancement to record linkage that we present.