Record Linkage in an Integrated Census

Theodor Yitzkov and Harry Azaria

Israel Central Bureau of Statistics
66 Kanfey Nesharim St. Jerusalem 95464, Israel, theodori@cbs.gov.il, harrya@census.cbs.gov.il

1. Introduction

One of the most important projects at Israel's Central Bureau of Statistics is the transition from a traditional census to an Integrated Census (IC) where the source for population counts is obtained from administrative files. To correct for the coverage of the population in the administrative files, large-scale sampling is carried out [6]. First, an area sample covering about 20% of the population is drawn and enumerated much like a traditional census. This is the U-sample and it is used to estimate the undercoverage of the administrative file. Second, all the persons listed in the administrative file in the sampling cells that were selected in the U-sample make up the O-sample. This sample is used to estimate the overcoverage of the administrative file, which results from people not living in Israel for more than one year and those not living in the area where they are listed in the administrative file.

The IC is based on an Integrated Administrative File (IAF), which is made up of several administrative sources linked together. The U-Sample File (USF) includes all the persons that were captured in the area sample. After linking the USF to the IAF, the O-Sample File (OSF) is defined. It contains all people listed in the U-sample cells but were not enumerated. These people consist of the potential overcoverage of the IAF, and they are further investigated for their current place of residency. In a traditional census, which is usually evaluated by a Post Enumeration Survey (PES), undercoverage in population counts can be adjusted using the classical dual system estimation procedure. In the IC, the dual system estimation procedure is extended to take into account the overcoverage as well. The estimation procedure is thus based on matching between the IAF, USF and OSF.

File matching relies heavily on the record linkage (RL) procedure. Therefore, it is necessary to develop reliable RL methodologies to ensure accurate census estimates for the population counts. In an IC, RL procedures are used intensively, including building the IAF from different administrative sources, linking the IAF and the USF to define the OSF, linking back the OSF after its investigation, and linking institutionalized persons and other special populations.

Two main approaches of RL are considered in this paper: exact matching and probability matching [3, 8]. We propose that the exact matching stage include matching not only by identification number, but also on sets of variables that are identified as having a pre-defined level of precision. This makes it possible to apply exact matching on a wider scale and to benefit from its advantages on large data arrays. It is also argued that the determination of precise variable sets for the exact matching stage solves the problem of defining the matching variables for the probability stage. The advantage of "early" definition of the matching variables is that it is based on the most complete existing file with respect to the coverage of the population. For probability matching, we develop a modified Jaro string comparator index to be compatible to Hebrew. We also propose a method for comparing the sensitivity of the agreement weights that are based on different matching variables. Using a regression model, we estimate the coefficients for each of the agreement weights for the final index. The estimates are based on the discriminating power of the matching variables. All results are illustrated for data from the first experiment of the IC carried out in one of Israel's towns.

2. The basic problems of record linkage in an Integrated Census

We can separate the RL problems in an IC into two categories: the traditional RL problems and those that are specific to the IC. The traditional RL problems result from the fact that the files belong to different databases. The files may have the following problems:

- The information is not complete or does not exist for some records;
- The information contains mistakes and distortions;
- A key variable, such as an Identification Number, may be missing or erroneous for some or all of the records;
- There are not enough variables that can serve as matching variables;
- The variables may have to undergo transformations to a common form before linking.

Insufficient matching information and erroneous data can result in errors such as false links, duplicate links (one to many or even many to many), and lack of links.

Besides the traditional RL problems there are specific problems related to the IC:

- The impact of false links on the reliability of the estimates is much more severe than that of a lack of links.
- It is necessary to perform the RL between the IAF and USF under a very limited time constraint in order to begin investigating the O-sample as soon as possible.
- There is no matching variable that can be used to divide the population into small blocks (e.g. households), which will make the RL procedure run quickly and correctly.
- The addresses in the IAF may be inconsistent with the actual addresses found in the USF, including synonym addresses, addresses not coded, and missing data on addresses.

Hence, the challenge of RL methods is to supply maximum true matches and minimum false ones between records from different sources. In the following sections we present improvements in RL methodology that address the problems listed above.

3. Exact matching

Exact matching is the RL procedure that uses key variables, and a match means that they are exactly equal [3, 8]. For example, exact matching between two (or more) sources by matching on the Identification Number (IDN) assures that the linked record pertains to the same person. We shall use this concept in a more broad sense: linking records that coincide on certain sets of variables that are not necessarily key ones and do not always identify a person.

For Israel's data on individuals, the role of exact matching procedures is particularly important for RL. First, because every citizen in Israel has an IDN made up of 8 digits and a check digit. When recorded, and recorded correctly, it is a powerful matching variable. Second, the probability matching may be more problematic than in usual applications because of the absence of reliable grouping variables. Hence, it can demand more time for calculations and for clerical review.

To verify the success of the link according to the IDN number, the first name of the persons is also checked. This is important since the IDN is usually reported by one of the household members for all members, and thus may be somewhat mixed-up. If the IDN exists but is erroneous (i.e. the check digit is incorrect), one can try to correct it. In this case, it loses most of its power after the correction, and has to be used with other matching variables. For this case and particularly for the case where IDN are not included in the files, we devote the following sections.

3.1 Precision of matching variables sets

Exact matching is carried out on a set of variables. If the set is complete enough to identify each person in the population, its precision is maximal. The question is how to choose and use matching variables for linking arbitrary files that may be of poor quality and how to estimate the precision of sets of matching variables that do not exactly identify a person.

Assume that there exists a Master File *C*, which covers almost all the population with variables of good quality, for example, the Central Population Register (CPR). The files *A* and *B* to be linked pertain to the same population, but may be of arbitrary quality.

In the following propositions we suggest that rather than link *A* and *B* directly, they should be linked through the high-quality Master File, and that the accuracy of this procedure is similar to that of identifying persons in the Master File. Figure 1 shows how files *A* and *B* can be linked through the file *C*. We further argue that given that it is sufficient to study a linkage to a Master File to assess the accuracy of any linkage, our second proposition estimates the probability of false matches for a given key *X*.

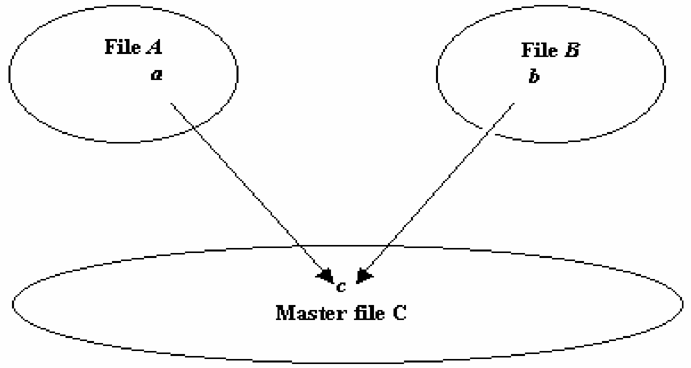

Figure 1. Schematic plot of linkage of files A and B through a Master File C.

Let a and b be persons in files A and B, respectively, and r_a a record representing person a in file A. The expression "$a = b$" means that records r_a and r_b represent the same person. Denote by X the set of matching variables. For convenience we assume that the finite set of all possible values of X is indexed by j. Let $X(a)$ the values of X for record r_a , K_j the set of all persons in file C with $X(c) = j$; $k_j = |K_j|$ the number of persons in the set K_j; and N_U the population size (since C contains almost all people in the population N_U is about the size of C). Finally, let P_j be the probability that a person a with $X(a)=j$ actually belongs to K_j, $P_j = P(a \in K_j \mid X(a) = j)$.

Proposition 1. Suppose that files A and B are two independent random samples from the population. Then the probability for a correct link between persons $a \in A$ and $b \in B$ with $X(a)=X(b)=j$ through a person $c \in C$ equals to $P_j^2/k_j + (1- P_j)^2 / (N_U - k_j)$.

Proof:
Because the files are a random sample from the population, then for any person a in A (or B):
$\forall j, \forall c \in K_j, P(a = c \mid X(a) = j) = P(a = c \mid X(a) = j, a \in K_j)P(a \in K_j \mid X(a) = j) = P_j / k_j$.

We define the events $L=\{X(a)=X(b)=j\}$, $L_a=\{X(a)= j \}$ and $L_b=\{X(b)= j\}$.

$$
\begin{aligned}
&P(a = b \mid L) = P(a = b \cap (a, b \in K_j) \mid L) + P(a = b \cap (a \notin K_j \cup b \notin K_j) \mid L) \\
&= \sum_{c \in K_j} P(a = c \cap b = c \mid L) + P(a = b \cap a \notin K_j \cap b \notin K_j \mid L) \\
&= \sum_{c \in K_j} P(a = c \mid L_a) P(b = c \mid L_b) + \sum_{c \notin K_j} P(a = c \cap b = c \mid L) \\
&= \sum_{c \in K_j} \frac{P_j}{k_j} \cdot \frac{P_j}{k_j} + \sum_{c \notin K} P(a = c \mid L_a) P(b = c \mid L_b) \\
&= \frac{P_j^{\,2}}{k_j} + \sum_{c \notin K_j} \frac{(1 - P_j)^2}{(N_U - k_j)^2} = \frac{P_j^{\,2}}{k_j} + \frac{(1 - P_j)^2}{N_U - k_j}
\end{aligned}
$$

The result can be interpreted in the following way. As file C becomes closer to the full population and P_j approaches 1, the probability to link correctly two persons with the same value of the matching variables $X = j$ through the file C is approximately $1/k_j$. On the other hand, the probability that any person with $X = j$ is placed in the group K_j and is linked correctly equals P_j/k_j, which also approaches $1/k_j$, as C becomes more complete and correct. This result leads us a) to perform all RL procedures through a Master File, and b) to use only the Master File for deciding which sets of variables to use for RL.

Proposition 2. The average probability P_X of a false match for any person in the Master file, based on the set of matching variables X, equals $(N_X - M_X)/ N_X$, where N_X is the number of records in the Master File that don't contain missing values in variables X, and M_X the number of different values of X.

Proof:
Using the complete probability formula, we write for any persons a and c in the Master File C:

$$
\begin{aligned}
P_X &= P(a \neq c \cap X(a) = X(c)) = \sum_j P(a \neq c \cap X(a) = X(c) \mid X(a) = j) P(X(a) = j) \\
&= \sum_j \frac{k_j - 1}{k_j} \frac{k_j}{N_X} = \frac{1}{N_X} \sum_j (k_j - 1) = \frac{N_X - M_X}{N_X}
\end{aligned}
$$

This average probability of a false match by variables X is used for estimating its precision. The set of variables X for which this probability is less than some threshold probability P_0 will be called a *sufficiently precise set of variables* (SPSV). That is, a set X that is SPSV satisfies $P_X < P_0$.

The conception of a Master File is discussed in many works. It is the file that contains a person IDN and maximal information about persons in the population. Only sufficiently complete files may serve in this role. In the IC, the first version of a Master File is the Central Population Register (CPR) that contains geographic and demographic information for all citizens of the country.

Table 1 contains the average probabilities P_X of a false match for various sets of variables X, calculated on the CPR for the first IC experiment, comprising about 50,000 residents. From the table we can conclude that variable set #4 may be considered SPSV. The variable Gender adds almost no precision to set #5. The Address variable (Street Code) adds considerable precision (sets #6 and #7), but in our case is not a stable one.

Table 1. Precision of matching variables sets X for the first IC experiment

Matching variables	Matching variable sets (X)						
	1	2	3	4	5	6	7
Last name	V		V	V	V	V	V
First name		V	V	V	V	V	V
Year of birth				V	V		V
Gender					V		
Street code						V	V
Average probability of a false match (P_X)	0.857	0.745	0.060	0.0035	0.0034	0.0036	0.00027

3.2 Use of matching variable sets

The SPSV can be partially ordered by the inclusion operator. It is clear that the precision of subsets of a given SPSV is lower than that of the full set of matching variables. Note that using a more precise SPSV gives fewer matches because of mistakes and missing values when adding more variables. Using a less precise SPSV, however, can result in more duplicate matches. Note that if the threshold P_0 is relatively low, then all duplicate matches are regarded as true ones. In this case it is recommended that only the less precise (or minimal) SPSV be used. A SPSV is *minimal* if removal of any variable from it results in a set that is not sufficiently precise.

There may be several minimal SPSV because the ordering is only partial. The minimal SPSV may be found in several ways. The algorithm that can provide most of them is the following:

1. Compose a full or a maximal set of matching variables (X_{max}).
2. Calculate the precision of each variable from X_{max} as if they were a separate set.
3. Remove the less precise variable and calculate the precision of X_{max} without the variable.
4. Repeat by removing variables one after another as it is described in 3 as far as P_X is less than the threshold P_0. The last SPSV is minimal. Call the resulting set of iteration $i - X^i$.
5. To get a new minimal SPSV repeat the process of items 3 and 4, starting with the removal from X_{max} of the least precise variable in X^i. The process is stopped when variables in all minimal sets $\{X^i\}$ are used in step 3.

The application of SPSV for exact matching may have some difficulties. There may be too few matching variables in the files, and as a result the number of SPSV will be too small. In this case, the researcher is compelled to enlarge theP_0 and to allow less precise sets. The recommendation in such situations is to first use the more precise SPSV, then the less precise ones. This way, on the first stage of the RL, one can link records that may have duplicates on less precise SPSV. For example, if the Master File itself is one of the linking files, a variable that has many missing values in the second file may be used in the first stage and then removed from the SPSV. The new set of matching variables (that is not sufficiently precise) may be used to get new links for records that have no duplicates.

3.3 Processing of duplicates

The problem of duplicates on matching variables may become the "nightmare" of the RL process. One-to-many, many-to-one and many-to-many matches (Master File-to-another file), which are carried out with not enough precision may complicate the results of the linkage, and make the analysis very difficult. Many difficulties may be avoided if the Master File is enriched while the additional files are linked to it.

Enriching the Master File. A Master File does not contain unknown duplicate persons, since the IDN marks all different persons. If a record from a new file matches to one of the records of existing persons in the Master File, then it is linked to that person. If there is no record in the Master File that matches to the new one, then a new person is created in the Master

File. Such a process enriches the Master File with persons and/or with demographic and geographic information. In the IC, we start by building the Integrated Administrative File by linking the CPR with additional files (local municipality file and data on students in the education system). Here the CPR is the Master File. We have decided that non-matched records will not be added to it since the CPR contains all Israeli citizens. However, we do enrich the CPR with geographic information. This decision ensures that duplicate persons are not created in the IAF.

Many-to-many and many-to-one duplicates. In the process of building the IAF, duplicates may occur on both of the files that take part in the RL procedure. Because the Master File does not contain duplicate persons, only one of its matches may be true. Therefore we do not allow links of the kind “many-to-many” or “many-to-one” between the Master File and a second file. This decision is consistent with the decision not to add new persons to the Master File.

One-to-many duplicates. Consider the duplicates of the kind one-to-many. Even if the set of variables is a precise one, the question of what to do with duplicates is not defined. The file to be linked may contain only different persons. In this case there is only one true match from among the duplicates. If the file to be linked contains duplicate persons, all the duplicates could be linked and produce true matches. The last case takes place in the RL procedure between the IAF and the USF. This is the second stage in the IC process and here the IAF is the Master File. The USF may contain duplicates that were recorded in different addresses. The correct address is chosen after the RL procedure. So all different addresses have to be collected first and linked to the one corresponding person. On the other hand, it is likely that not all such links are correct, as may be the case for any non-duplicate link. The expected number of false links is defined by the precision of the SPSV.

Choosing between alternative values. Different values of a certain variable may be attached to the same person. The question is, which value has to be used for the matching. In practice, the problem has too many solutions and the answer depends on the stage of the RL. In our experiment, when building the IAF all variables that exist in the CPR, except for the address variable, are considered as more correct and are used as IAF variables. The address is the only variable that may be used in different variations. The algorithm of choosing the correct address from among all possible addresses in the different administrative files is complicated and will not be described here. The end result is that the IAF contains only one address variable.

4. Probability matching

The theory base for probability matching was developed by Fellegi and Sunter [2]. Two kinds of errors are examined: links of records that belong to different entities (type I error or false matches) and missing the records that belong to the same entity (type II error or false unmatches). Due to the Fellegi and Sunter theorem, the optimal likelihood function for the classification of record pairs is

$$R = \frac{P(\gamma \mid (a_i, b_j) \in m)}{P(\gamma \mid (a_i, b_j) \in u)} = \frac{P_1}{P_2}$$

where a_i and b_j are entities from the linked files A and B, m is the set of all true matches, u the set of all false matches, and $\gamma = (\gamma_1, \ldots, \gamma_l)$ a vector that represents a specific comparison on matching variables. The rule for the classification of record pairs

$$Z(a_i, b_j) = \begin{cases} \text{matched} & \text{if } \; R > R_+ \\ \text{demand review} & \text{if } R_- \le R \le R_+ \\ \text{nonmatched} & \text{if } R < R_- \end{cases}$$

minimizes the number of record pairs that demand clerical review for given levels of errors of both types. For calculating the numerator P$_1$ it is usually assumed that the matching variables are independent (although this is a strong assumption) and therefore $P(\gamma \mid m) = P(\gamma_1 \mid m) \ldots P(\gamma_l \mid m)$. Similar argument is used for P_2. Hence the statistic is of the form

$$W_{opt} = \prod_{i=1}^{l} \frac{P(\gamma_i \mid m)}{P(\gamma_i \mid u)}$$

In our practical calculations we use the linear form to get the overall agreement weight

$$W = \sum_{i=1}^{l} \beta_i W_i \qquad (1)$$

where W_i is an *agreement weight* of a matching variable *i* and the coefficient β_i is called the *discriminating power coefficient*. We show now that there is a similarity between the elements of W_{opt} and W, noting that the first statistic is multiplicative while the second is additive.

The denominator P_2 of R is the proportion of record pairs that have a similar level γ among all false matches. The number of all false matches $|u|$ is large and has the order of $N_A N_B$. The nominator of P_2 is usually much smaller than $|u|$. Hence, changes in the numerator will cause a small change in proportion P_2. So instead of P_2 that is dependent on a specific value of γ, we replace it with P_{ui} - the ratio of record pairs with the same values of matching variables among all false matches. The value of $1/P_{ui}$ is larger for variables that have more distinct values, so it characterizes the discriminating power of a variable i. The transition from the ratio P_2, that is specific for a particular agreement level, to P_{ui}, that reflects the level of complete agreement, is similar to the transition from a specific frequency ratio (for a specific variable value) to a global frequency ratio (for the all possible variable values) [3].

Hence, going back to W, we note that W_i has the same meaning as $P(\gamma_i \mid m)$, and can get values between 0 and 1. The coefficients β_i have the same meaning as $1/P_{ui}$. The probability of a true/false match is calculated for the overall agreement weight W.

4.1 Adjusting Jaro's string comparator index for Hebrew

We begin by describing an agreement index for alphanumeric variables. The Jaro string comparator index compares empirically two strings, and is of the form

$$J_1 = 0.5(common / Len1 + common / Len2),$$

where $Len1$, $Len2$ are lengths of the strings, and $common$ is the number of letters in common in both strings.

Different variants of the Jaro index have been considered [3-5,7-8]. A popular form of it uses instead of common letters, the number of similar letters (*similar*) and the number of transpositions of similar letters (*transpos*), where the letters reside in the same half of the string:

$$J_2 = \frac{1}{3}\{similar / Len1 + similar / Len2 + (1 - 0.5 \cdot transpos / similar)\}.$$

It is clear that the concept of similar letters depends on the alphabet and on the language. The similarity of letters may be a consequence of two reasons: typical typographical errors between similar letters and the way the letters are pronounced. We have composed a table for similar Hebrew letters for these two reasons and use this table for the above calculation. Searching for similar letters is carried out for letters within distance that is less or equal to (*Len1* + *Len2*) / 4 (arithmetic average of the string's half length). The total number of similar letters is calculated as a weighted average where common letters receive a weight of one and the similar letters according to the table receive a weight of 0.5.

Instead of the number of transpositions we study the total distance between similar letters (DistTot). In most cases, we obtain the inequality: $DistTot \leq (LenM - 1)(0.5\ LenM - 1) = DistMax$,where $LenM = max(Len1, Len2)$. Hence, the term $(1 - DistTot / DistMax)$ may be used as a measure of how correct the placements of the letters are. This term replaces the third term in expression for J_2. In the rare case where the above inequality does not hold, this term can be a negative, although in practice the index is positive.

Our experience leads us to make necessary changes to minimize the influence of this term for the case when there are not many similar letters in the strings. If there are no similar letters in the strings (or only one and in the right placement), the DistTot (and *transpos*) is equal to 0, and so the third component in J_2 is equal to the maximum value of 1. We believe this overestimates the level of similarity and gives a false high value to strings that are not likely to be a match. The proposed form of the modified Jaro Index is therefore

$$J_M = \frac{1}{3}[Similar / Len1 + Similar / Len2 + (\frac{Similar}{LenM})^2 \{1 - \frac{DistTot}{(LenM - 1)(LenM / 2 - 1)}\}].$$

The calculation of the Jaro Index is in the last stage of the string processing. At the first stage, strings go through some preliminary processing by removing notations, hyphens, etc. There is also a common practice in Israel to give two first names. These are written as two parts of one string in the same data base field. One of the preliminary tests checks matches with each part separately. If there is a match, the J_M gets a value 1. Other specific processing is carried out for problems relating to the language.

4.2 Agreement weights for age and gender variables

The age variables may be very precise if they contain the year, month and day of birth. However, if the month and day variables are not of high quality they may not be used for exact matching and will not add much to the probability matching either. This turned out to be the case for the USF data where the day and month of birth were not reliable and had many missing values. Therefore we did not use them in the exact matching stage. The typical error in the year of birth was a difference of 1-3 years. It was decided to use the following function on the difference between years of birth, which quickly decreases, as the difference gets larger:

$$W_{age} = Exp\left(-\,|\,Age1 - Age2\,|\,/\,k\right) \; .$$

The parameter k may be estimated in the process of RL, but we used the initial value k = 3. In this work we did not set more weight to ten years difference (a typical typographical error) or for similar figures (8 – 3 – 5, 1 – 4 – 7 – 9).

For the calculation of an agreement weight for gender we used the following simple function:

$$W_{gen} = \begin{cases} 1 & \text{if} \quad Gender1 = Gender2 \\ 0 & \text{if} \quad Gender1 \neq Gender2 \\ 0.5 & \text{if} \quad Gender1 \quad \text{or}\, Gender2 \,\text{is missing} \end{cases}$$

The variable gender is not significant in probability matching as it was shown to be not significant in the exact matching stage.

4.3 Estimation of discriminating power coefficients

To use formula (1) it is necessary to know the coefficients β_i that reflect the discriminating power of the matching variables X_i. In this work, we start with ad-hoc expert estimates for the coefficients of the matching variables: for First Name the coefficient is equal to 10, for Last Name it is 10, for Year of Birth 6, and for Gender we use the value 4. The IAF and USF for our data were linked according to the maximal value of the overall agreement weight W. The resulting file of record pairs was subjected to complete clerical review. Each record pair received a value Z= 1 if the link was confirmed and Z = 0 if it was rejected. The variable Z was used as the dependent variable in a logistic regression model of the form:

$$Z = \text{logit}(W) = \frac{\exp(W)}{1 + \exp(W)}, \quad W = \beta_0 + \sum_{j=1}^{4} \beta_j W_j$$

Table 2 shows estimates of the discriminating power coefficients, the coefficients after calibration (setting the Last Name coefficient to 10 and retaining proportions between coefficients) and quality of regressions by the C-statistic (rank correlation index). The first two rows describe results of the first stage of RL using a geographical blocking variable. The third row describes RL of residuals after the first stage using a different blocking variable (first letter of the last name). The results show that the coefficients for Last Name, First Name and Year of Birth are highly statistically significant and rather stable for all stages. The coefficient for Gender is not significant and even once had a negative value.

Table 2. Estimation of the discriminating power coefficients for RL between IAF and USF in the first IC experiment

Stage of RL	Coefficient / Calibrated coefficient P-value				C - statistic
	Last name	First name	Year of Birth	Gender	
Preliminary IAF	12 / 10 < 0.0001	7.9 / 6.6 < 0.0001	3.0 / 2.5 < 0.0001	-0.30 / -0.25 0.66	0.99
Final IAF	11 / 10 < 0.0001	6.8 / 6.2 < 0.0001	2.5 / 2.3 < 0.0001	0.30 / 0.27 0.54	0.99
Final IAF using a different blocking variables	7.2 / 10 < 0.0001	4.8 / 6.7 < 0.0001	0.7 / 1.0 < 0.04	0.20 / 0.28 0.59	0.90

4.4 Estimation of thresholds for automatic classification of matches

Complete clerical review cannot be carried out for large data processing. Hence we must use more limited resources to achieve the aim of RL, formulated at the end of Section 2. In practice, most of the matches between record pairs must be classified automatically. There are two approaches for classifying matches: direct and indirect [1]. According to the direct approach, the distribution $g(Z_i \mid W_i, \nu)$ is estimated (ν -a vector of parameters). This approach may be typified by logistic regression $P_i = \text{logit}(\nu_0 + \nu_1 W_i)$. The indirect approach is estimating the distribution $f(W_i \mid Z_i, \phi)[\lambda^{Z_i}(1-\lambda)^{1-Z_i}]$ (ϕ - a vector of parameters, λ - the probability that the match belongs to the true matches class). The indirect approach is typified by discriminant analysis and hypothesis testing methods. We proceed with the indirect approach.

The statements that have to be checked are *Z(a, b)* = 0, that is, *a* and *b* are different persons against *Z(a, b)* = 1, *a* and *b* are the same person. To find the upper threshold W_+ for which all the record pairs that satisfy $W(a, b) > W_+$ are considered as true matches, it is necessary to set: H_{01}: *Z(a, b)* = 0 and H_{11}: *Z(a, b)* = 1. The lower threshold W_-, for which record pairs that $W(a, b) < W_-$ are considered as false matches, can be found by exchanging the roles of the null and alternative hypotheses, naming them H_{02} and H_{12}, respectively. Clerical review must be carried out in the interval $[W_-, W_+]$. All the efforts of improving the agreement weights are intended to raise the sensitivity of the statistical criteria and to narrow the clerical review interval. Figure 2 shows RL results of IAF and USF for our data. The true and false matches were defined by complete clerical review of approximately 1200 matched record pairs.

The Y-axis on Figure 2 shows the probability and the X-axis the values of the agreement weight. The curves represent F1, the empirical c.d.f. of true matches ($F(W \mid Z = 1)$), F0, the empirical c.d.f. of false matches ($F(W \mid Z = 0)$), and P, the predicted values of the logistic model of Z on W.

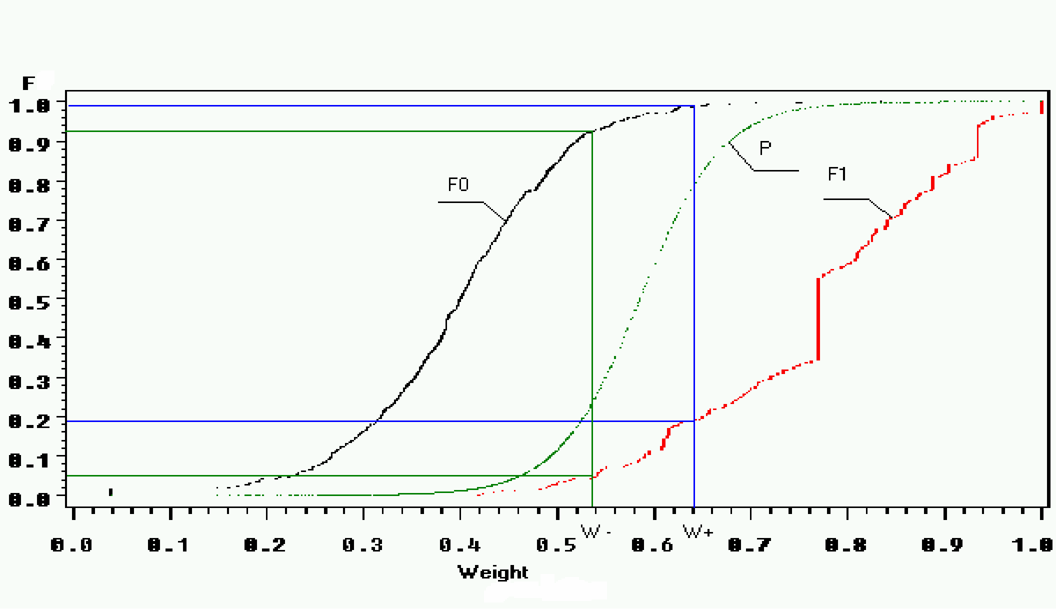

Figure 2. Empirical c.d.f. of agreement weight W. The curves represent F1, the empirical c.d.f. of true matches ($F(W \mid Z = 1)$), F0, the empirical c.d.f. of false matches ($F(W \mid Z = 0)$), and P, the predicted values of the logistic model of Z on W .

For testing hypothesis H_{01} the type I error α_0 was set to 0.01, and for testing H_{02} the type I error α_1 was set to 0.05, because for the IC it is more important not to get a false link than to miss a true match (see Section 2). The discriminating capability of a test is estimated by its power. For the first test we get a power of $1\text{-}\beta_1 = 0.8$, and for the second test $1\text{-}\beta_0 = 0.93$ (the power value depends on the level of the type I error). The thresholds are W_- = 0.54, W_+ = 0.65. The estimate of probability λ (percent of correctly linked records) is 0.256 and clerical workload may be estimated by $h = (1-\lambda)(\beta_0 - \alpha_0) + \lambda(\beta_1 - \alpha_1) = 0.083$.

The same calculations were done using the basic form of the Jaro index J_2 with *common* number instead of *similar* (because similar Hebrew letters were introduced in this work). In this case the power was 0.5 for the first test and 0.7 for the second. The thresholds are W_- = 0.6, W_+ = 0.8. The estimate for clerical workload *h* = 0.429 indicates the advantage of the modified Jaro index.

5. Summary and discussion.

RL starts with exact matching and continues to probability matching. The question, which variables to use for matching, usually arises in the probability matching stage. The suggested approach in this paper is to define sufficiently precise sets of variables already in the exact matching stage. Such sets of variables may be used for exact matching with little likelihood of mistake. Sequential use of exact matching by different sets of variables enables finding most of the correct matches in a short time even for large files. We also suggest that any two files, which need linking, will be linked through a master file. The advantages of our approach are:

- The estimates of the variables' precision are built on the basis of an existing file (e.g., CPR) and the decision rules are obtained before the RL process.
- By using SPSV the exact matching may be carried out quickly without the complicated procedures of examining record pairs and clerical review.
- The information about which sets of variables and in what order we should use them is required anyway for probability matching.
- The concept of RL using a Master File can help avoid some of the problems due to duplicates and the matching precision will make the RL more reliable in due course.

The difficulties that result from using SPSV in the exact matching stage are related to the processing of duplicates, the possibility of erroneous links, and how to process the duplicates. These are inherent problems in probability matching, and it has been shown here that they can be solved in the exact matching stage as well.

After exact matching is carried out, the record linkage is carried out by probabilistic methods. For linking names, an innovative agreement measure is calculated based on a modified Jaro string comparator index. The modified index is sensitive to differences in the strings and more adapted to Hebrew. For our data, the index used for determining the matching status is calculated as a weighted sum of agreement indices for names, age and gender. The weights are determined by a logistic regression model. The model is built for data sets that passed clerical review and had the true matching status for each record. The upper and lower thresholds for determining the different classes of the matching status are based on the classical Neyman-Pearson method for hypothesis testing. The methods are illustrated and tested on real data from the first experiment of the Integrated Census in one town in Israel. The proposed methods enable us to get preliminary estimates of the parameters (weights, thresholds) that can be used as first approximation for the next IC experiment. A major problem in our data is the lack of reliable blocking variables (such as households). This problem may cause lengthy matching procedures and hence may restrict usage of probability methods for determining the OSF.

6. Bibliography

[1] Belin T R. and Rubin D B. (1995) Method for calibrating false-match rates in record linkage. *Journal of American Statistical Association,* **90**, 694-707.

[2] Fellegi I P and Sunter A B. (1969) A theory of record linkage. *Journal of American Statistical Association*, **64**, 1183-1210.

[3] Gill L E. (2001) *Methods for automatic record matching and linking and their use in national statistics.* National Statistics Methodological Series, Great Britain, Oxford University, Number 25.

[4] Jaro M A. (1985) Current record linkage research. In *Record linkage techniques, U.S. Internal Revenue Service*, 317-320.

[5] Jaro M A. (1995) Probabilistic linkage of large public health data files. *Statistic in Medicine,* **14**, 491-498.

[6] Nirel, R., Glickman, H. and Ben Hur, D.(2003) A strategy for a system of coverage samples for an integrated census. Unpublished manuscript.

[7] Winkler W E. (1988) Using EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 667-671.

[8] Winkler W E. (1994) Advanced methods for record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472.

Acknowledgements

The authors thank to Ronit Nirel, Hagit Glickman, Natalie Shlomo and Dan Ben Hur for their help on all stages of the work and helpful comments on previous versions of this paper.