# Testing Disclosure Risk in the proposed SIPP-IRS-SSA Public Use Files

John Abowd, Sam Hawala, Bryan Ricchetti, Martha Stinson

November 16, 2006

## 1 Overview

As the result of a four year joint project between the Census Bureau, the Internal Revenue Service, and the Social Security Administration, the LEHD Program has created an enhanced SIPP file that links a subset of SIPP variables to administrative earnings and benefits data. We have reviewed this file for disclosure risk and here present our results to the Census Disclosure Review Board. We believe that the procedures we used to create the synthetic data conform to the Census Bureau's disclosure avoidance requirements and request that the DRB grant permission for the file release. We understand that the disclosure officers at SSA and IRS must also certify that the file meets their agency's confidentiality requirements. We have requested such releases from SSA and IRS, both of which requested that Census first determine whether the file meets the Title 13 requirements for release.

The link between administrative earnings, benefits data and SIPP data adds a significant amount of information to an already very detailed survey (for details on the creation of this linked file, see "Final Report to the Social Security Administration on SIPP/IRS/SSA Public Use File Project") and could pose potential disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The creation of synthetic data is meant to prevent a link between these new public use files and the original SIPP public use files, which are already in the public domain. In addition, the synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information.[1] Our disclosure avoidance research uses the principle that a potential intruder would first try to re-identify the source record for a given synthetic data observation in the existing SIPP public use files, which were used to create the SIPP component of our Gold Standard file.

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted two distinct matching exercises between the synthetic data

[1] A formal request to the IRS disclosure officer to confirm this statement with an official memo has been made.

and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, the Gold Standard variables are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the Gold Standard and the synthetic data represent potential disclosure risks.

It is important to remember that for an actual re-identification of any of the records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required. This additional step consists of making another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc.* Hence, the results from our matching process are a very conservative estimation of re-identification risk.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. Second, the ratio of true matches to the total number of matches (true and false) should be consistent with considerable uncertainty regarding which matches are "true" and which are "false."[2] We have performed two types of matching exercises, probabilistic and distance-based. This section describes the results from both exercises and gives an assessment of the risk of disclosure associated with the synthetic data files.

# 2 Matching based on probabilistic record linking

We begin with the probabilistic record linking experiment. Since the public use files consist of 16 different implicates, one must consider the risk associated with each file. In previous runs of this matching process, similar results were found on the different implicates. The evaluation of disclosure risk described in this section centers on the risk presented by the publication of one single implicate file (the first synthetic implicate that matches to the first missing data implicate, *i.e.* $m = 1$ and $r = 1$). In view of the results that are described below, we expect that similar results would be obtained for the other implicate files individually. In section 3 we will evaluate the disclosure risk presented by

[2] Originally, the DRB proposed that the ratio of true matches to false matches should be about 1.0, indicating that the "best match" was about equally likely to be true as false. We do not think this standard is appropriate for the current re-identification exercise because the intruder knows that the source of every record in the proposed PUF is in the Gold Standard file. Hence, the optimal matching strategy declares the record with the highest match score to be a match, regardless of the agreement score. Hence the ratio of the true match rate to the false match rate in our analyses is always about 0.03–which appears to easily meet the DRB criterion. Hence, we adopted a more conservative criterion: we compare the success rate of the best match to the success rates of the second and third best matches. The best match outperforms the second and third best matches, as it should, but not by very much. This means that a potential intruder would face considerable uncertainty regarding choice among the top three match candidates as to which one is the true match.

the file obtained by averaging the variables across all the implicate files.

Probabilistic matching requires creating a set of blocking and matching variables that are common to both files. We implemented one blocking strategy using the unsynthesized variables for blocking. For married individuals we use the unsynthesized variable *male* for each member of the couples. For unmarried individuals we use the two unsynthesized variables, *male* and *maritalstat*. The latter can be either widowed, divorced/separated, or never married ($maritalstat = \{2, 3, 4\}$). In other words, for two records to be a match, they must necessarily have identical values for marital status and gender since these two variables were not synthesized. After this has been determined to be the case, other variables can be compared to determine the probability that two records represent the same person.

The probabilistic record linking was performed using the Census Bureau's internal record linking software, which is maintained by the Statistical Research Division. The discussion in this section describes the technical settings used for that software. We set the blank filter flag equal to 0 so that if the variable is missing, the record will automatically be considered to agree on that field. Matching for the two groups, married and unmarried, was done separately. Blocking variables help to reduce the number of records used for comparison; however, in any given run all records in the same blocking group of the synthetic implicate and the Gold Standard files are compared. Thus, record linking computation is quadratic with run times dominated by the size of the largest block. In this latest version of the SIPP/SSA/IRS-PUF, the block sizes are very large. For this reason, the matching is done within corresponding segments of the Gold Standard and PUF files. Internally we know when segments of the Gold Standard and PUF files (single implicate) correspond to the same individuals, because we make use of the common artificial person identifier (*personid*) that is on both files. Without the information contained in *personid* (which is not on the actual PUF), an intruder would have to compare many more record pairs to find true matches and would not find any more true matches (the true match is guaranteed to be in the blocks being compared) and would almost certainly find more false matches. For this reason our approach leads to a conservative measure of the disclosure risk.

When the SIPP/SSA/IRS-PUF is finally publicly released there will be no link between the Gold Standard data and the synthetic implicate files. However for testing purposes, we have maintained this link by keeping the common person identifier on the Gold Standard file and the PUF implicate files. Thus, by naming this person identifier in the sequence field of the record linking software, we can check which matched record pairs with a given score are correct matches and which are false matches using this person identifier. When the person identifier is the same, the matching algorithm was successful in finding the person in the Gold Standard file from whom the synthetic data record was generated. When the person identifier is different, the matching algorithm was unsuccessful. This technology is also used for the distance matching discussed in section 3.

Automatic searches for matches occur only within those records sharing the

same values on the blocking variables. Matches agree exactly on values for the blocking variables and, additionally, they agree on values for the matching variables. An input file to the matching software specifies the agreement criterion for each of the matching variables. Two numbers have to be specified for each of the matching variables. The first number represents the conditional probability that the two records agree on the matching field value given that the two records represent a match, called the $m$ probability. The second number represents the conditional probability that the two records agree on the matching field value given that the two records do not represent a match, called the $u$ probability.

From the agreement criterion, the software computes a score. The agreement score for a match on a particular variable from two comparison records is based upon $\ln(m/u)$. A larger ratio implies a stronger distinguishing power for that matching field. Presumably, the ratio $m/u > 1$. When using Census Bureau matching software for the un-duplication of a file, one is trying to identify specific duplicate pairs, so more precise probability estimates may be helpful. However, when using this software for extracting subsets of plausible matches from a large file, the conditional agreement probabilities can be rough general estimates. To use a more aggressive assessment of disclosure risk, we obtained the best possible $m$ and $u$ estimates by using the *personid* variable that is common between the files even though the estimation of those probabilities requires knowledge of the link. We have enough confidence in this technology that we believe these $m$ and $u$ estimates should be public information and have performed the disclosure analysis on that basis. Since these are the best $m$ and $u$ estimates, an intruder trying to match the two files cannot possibly obtain better results using matching software that is at least as efficient as the Census Bureau software.

It is easy to calculate the conditional agreement probabilities $m = \Pr(agreement \mid match)$ for each matching field, if one knows when true matches occur. This is just the relative frequency of the fields on the Gold Standard and PUF files being equal, call this $f_0$. It is also easy to calculate the unconditional probability $\Pr(agreement)$ for each matching field that has a categorical variable. If, for example, $X$ is a categorical variable that can take on 3 possible values, $x_1$, $x_2$, $x_3$ then we obtain the distributions of $X$ in the Gold Standard (GS) and PUF files (implicate 1) and calculate

$$\Pr(agreement) = \sum_{i=1,2,3} \Pr(X = x_i \mid GS) \Pr(X = x_i \mid PUF).$$

Next it is clear that $\Pr(match) = \frac{1}{N}$, with $N$ being the common size of both the GS and the PUF files, since for each GS record there is only one PUF record representing the same person. Therefore $\Pr(nonmatch) = \frac{N-1}{N}$, so given $m = \Pr(agreement \mid match) = f_0$, we have

$$\Pr(agreement) = \frac{f_0}{N} + \frac{\Pr(Agreement \mid nonmatch)(N-1)}{N}$$

and can solve for $u = \Pr(Agreement \mid nonmatch)$.

The agreement and disagreement conditional probabilities for those variables used for matching are shown in Tables 1 and 2. Table 1 refers to individuals

with spouses and Table 2 refers to individuals without spouses. All matching fields except *birthdate* were assigned the "exact" matching comparison type "c," which makes the program assign full agreement/disagreement scores according to whether the fields agree or disagree. For *birthdate* the matching comparison type was "d", which makes the program assign full agreement weight if the difference between the birthdates is less than one year (365 days), otherwise the program assigns the following score:

$$\begin{aligned} & A + (D - A) \times \frac{DIFF}{MAXDIFF} \\ A &= Agreement\ Score \\ D &= Disagreement\ Score \\ DIFF &= difference\ between\ synthetic\ and\ original\ values \\ MAXDIFF &= \max imum\ difference \end{aligned}$$

These probabilities are used to calculate the scores given to this variable when it agrees or disagrees. The agreement score is defined as $\ln(\frac{m}{u})$. The disagreement score is defined as $\ln(\frac{1-m}{1-u})$. For example, the full agreement score for a "c-match" on Hispanic is $\ln(\frac{0.888222038}{0.817697432}) \approx 0.08$. The disagreement score is $\ln(\frac{1-0.888222038}{1-0.817697432}) \approx -.50$.

The software compares each matching field, decides whether the field agrees or not, and then assigns the appropriate score to the field based on the user supplied $m$ and $u$ probabilities. Next, a cumulative match score is calculated by summing the scores across all the matching variables. This cumulative score is used to decide whether two records match. It is compared to the cutoff values provided by the user and if it passes the stated threshold, a match is declared. The influence of a one variable relative to another on this cumulative score is controlled by the relative matching and non-matching agreement probabilities specified by the user, but in this case based on actual calculations from the relevant files. The non-matching agreement probability essentially tells how often a field will agree at random across two files. A high value for this probability will reduce the importance of this variable in the matching by causing the agreement score to be lower. This is desirable because if the field is likely to agree at random, any match in values between two files is less likely to signify a true match. At the same time, a high non-matching agreement probability causes the disagreement score to be less negative or smaller, meaning that the penalty for not matching on this variable is not as high. In contrast, the relative matching agreement probability tells the importance of this variable compared to other variables in determining whether two records are a match. A high matching agreement probability means that a match on this field is crucial to determining an overall match between two records. Thus a high value for $m$ produces a high agreement score. It also produces a more negative or higher disagreement score, more severely penalizing non-matching in this field. Consider the example of the variable $flag\_mar4t$, which is used to identify individuals who reported more than three marriages. When two records agree on this variable,

and they are a match, the cumulative matching score increases by 5.317686217. If the records are not a match, but agree on this variable, then the cumulative score decreases by −4.609063992.

The output cutoff flag for the cumulative matching score provides the comparison points for the matching score. In our testing we declare any pair of records with a cumulative score between −20 and 20 to be a *potential* match. From either Table 1 or 2 we can see that the total matching scores cannot be outside of this range. Essentially, we allow every record in the synthetic file to have candidate matches in the Gold Standard. Most applications of probabilistic record linking use a positive cut-off for the automatic selection of potential matches. However applications with this feature are usually concerned with determining *whether* a record from one file has a matching record in another file. We know with certainty, and any intruder would also know with certainty, that every synthetic record has a match in the Gold Standard. Thus, from among the potential matches, we choose the three highest scoring matches for every synthetic record, even when some or all of the matching scores are negative, with the idea that we are choosing the best matches possible.

In the second column of Table 3, we report the number of true matches found among the highest ranking, second highest ranking, and third highest ranking matches, for both married and single individuals. Because every synthetic record was declared to have three matches, the number of false matches is (1-number of true matches) and does not provide any additional information. Instead, in the third column, we report percentage of total matches that were true matches. The highest scoring matches are true matches 3.28% of the time for married individuals and 3.21% of the time for single individuals. Among the second highest scoring matches, approximately 1% are true matches both for marrieds and singles, with almost identical rates for the third highest scoring matches.

In the fourth column of Table 3, we look at the ratio of true match rates for the second and first highest scoring matches. This statistic provides an odds ratio for how much more likely the first highest scoring match is to be true than the second highest scoring match. A ratio of one would mean that the first and second highest scoring matches were equally likely to be the true match. In these data, for married individuals, the ratio is approximately .4. When second and third ranking matches are pooled in column 5, the ratio becomes .75.

# 3 Distance matching

Distance-based record linking is another common approach to estimating the risk of disclosure in micro data. In recent work, [**?**] use distance-based methods to re-identify records on two synthetic micro-data samples. They find that distance-based metrics perform similarly to (if not better than) the more commonly used probabilistic methods. Their work suggests that re-identification exercises should also include distance based methods because, unlike probabilistic record linking, the distance measures can take proper account of correlation

among the variables. The broader the selection of methods used, the more informed the analyst is of the risk of disclosure. In particular, it is important to understand which methods pose the largest threat. [**?**] conduct similar comparisons of distance-based and probabilistic record linking methods.

Our tests consider the case of an intruder who uses distance-based re-identification to match the source records from the Gold Standard to synthetic SIPP/SSA/IRS-PUF observations. Such re-identification methods calculate the distance between a given record in the Gold Standard and every record in the synthetic implicate. The $j$ closest records are then declared potential candidates for a match to the source record. In our analysis we consider $j = 3$.

Our distance-based re-identification proceeds in two stages. First we split both the Gold Standard and the first synthetic implicate ($m = 1$ and $r = 1$) into groups based on the unsynthesized variables. In this case, marital status and *male* are the only two unsynthesized variables. We next split each blocking group into smaller segments of approximately 10,000 observations in order to decrease the processing time, which is quadratic in the size of the largest files compared. We performed the segment split on both the Gold Standard and synthetic files so that the correct match in the Gold Standard was always in the same block and segment of the synthetic data used for comparison. In other words, we forced the segmentation of the files to guarantee that the correct match could always be found in the block/segments being compared. This is the same assumption as we used in section 2 to segment the comparison files in that analysis. The segmentation of the blocks uses our prior knowledge of which records are actual matches and hence our matching results are conservative–overestimates as compared to a distance record link that could not segment the comparison files because the intruder did not have access to the true *personid*. After splitting the data into blocking groups and segments, we then calculate the distance between a given Gold Standard record and every record in the synthetic file in its corresponding blocking group and segment using a set of 163 matching variables. This list of matching variables comprises every SIPP variable included in both the synthetic and Gold Standard data. The three closest records are then declared possible matches.

We use four distance metrics. Each metric is a special case of either Mahalanobis or Euclidian distance. Before formally defining the distance, we first define some notation. Let $A$ and $B$ represent the two data sets being matched. For our purposes, conceptualize the block and segment of the Gold Standard as the $A$ file and the block and segment of the synthetic implicate as the $B$ file. Denote $\alpha$ as the vector of 163 matching variables from an observation in the $A$ file and $\beta$ as the analogue for the $B$ file. Given this notation we define the distance between a given vector $\alpha$ in the $A$ file and a given vector $\beta$ in the $B$ file as follows:

$$d(\alpha, \beta) = (\alpha - \beta)\prime[Var(A) + Var(B) - 2Cov(A, B)]^{-1}(\alpha - \beta)$$

We consider four specific cases of the general distance. In the first case we assume that the intruder can properly calculate the $Cov(A, B)$. We denote this

distance $MAHA1$, and note that it is a true Mahalanobis distance; hence we expect that this distance measure will give us the highest match rates since it uses all of the available information, including the correct covariance structure of the errors in synthesizing all 163 variables. In the second case, we assume that the $Cov(A, B) = 0$. This is equivalent to assuming that we do not know how to link the observations across the $A$ and $B$ files and cannot compute $Cov(A, B)$. A real intruder would not have access to $Cov(A, B)$. We denote the second distance $MAHA2$, and note that it is a "feasible" Mahalanobis distance. In the third case, we assume $[Var(A) + Var(B) - 2Cov(A, B)] = I$, where $I$ is the identity matrix. We denote the third measure as $EUCL1$, which is a Euclidian distance with unstandardized inputs. For the fourth measure, we transform all of the matching variables in the $A$ and $B$ files to $N(0, 1)$ variables. Call the transformed files $\tilde{A}$ and $\tilde{B}$. We then calculate the distance using $[Var(\tilde{A}) + Var(\tilde{B}) - 2Cov(\tilde{A}, \tilde{B})] = I$. We denote this fourth metric $EUCL2$, and note that it is a standardized Euclidian distance.

Tables 4 and 5 show the results of the re-identification exercises for each of the four metrics. Table 4 shows the results using the Mahalanobis distance measures and Table 5 shows the results for the Euclidian distance measures. For each metric there are six columns. Match rate 1 (closest two records in $A$ and $B$), match rate 2 (second closest two records in $A$ and $B$), ratio of 2/1, match rate 3 (third closest two records in $A$ and $B$), ratio of 3/2, and ratio (3+2)/1. Match rate $j$ is calculated as the number of successful matches within a blocking group based on the $j$th closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages). For example, match rate 2 is calculated as the number of successful matches within a blocking group and segment based on the second closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages).

We first note that match rate 1 finds the highest rate of re-identifications. This implies that choosing the closest record using the indicated distance metric is more likely to find true match than choosing the second or third closest record. We further note that the highest match rate among all blocking groups is only 2.91%. Thus, an intruder who defined the closest- distance record as a match would correctly link 1.09% of records overall in the synthetic files and less than 3% in the worst-case sub-group.

The three ratio columns give us a sense of how much better the closest match does than the second and third best matches. Ideally, we want to ensure that if an intruder looked at the top three matches, he or she would face sufficient uncertainty about which one was the correct match. If the second closest record is exactly as likely to be the correct match as the closest record, then the ratio of match rate 2 to match rate 1 would be unity. If this ratio is less than one, then the closest record is more likely to be the correct match. If this ratio is greater than one, then the second closest record is more likely to be the correct match. The other ratio columns have the same interpretation. For the $MAHA1$ metric, the column Ratio (3+2)/1 ranges from 0.79 to 1.12. This suggests that the 2nd or 3rd closest matches are almost as likely to be correct as the closest match. The totals in the last row are essentially weighted averages

of each column where the weights are the percentage of records in each group.

As expected, the $MAHA1$ metric produces the highest match rates. The highest match rate for the $MAHA2$ metric, perhaps the most likely to be used by an intruder, is 2.2% and the ratio of (3+2)/1 is very close to unity for every sub-group. The Euclidian metrics are very similar to the $MAHA2$ metrics with the overall match rate not exceeding 1.2%, the highest sub-group match rate less than 2.4%, and the ratio of (3+2)/1 generally being very close to or slightly higher than unity.

After calculating these matching rates, we re-ran our exercise with several variations. First, we added the pooled weight created to accompany the public-release file to the list of matching variables. While this variable is not an original SIPP variable, it was created using public-use SIPP data and, in theory, could be re–produced by an intruder. Thus in order to mitigate disclosure risk, the weight was also synthesized. We ran a second matching exercise where we calculated the distance based on the 163 SIPP variables used previously, and additionally, the weight. The changes in our match rates were neglible. This is good news from two perspectives. First, the synthesis of the weight appears to have successfully protected the variable. Second, this result may imply that we could actually release the unsynthesized pooled weight as a new SIPP variable that could be used by researchers wishing to combine the five panels conducted in the 1990s. We are not requesting such a release at this point in time but think it will be worth considering in the future.

Finally, we averaged the 163 SIPP variables and the weight across the 16 implicates to create a new "average implicate" file. This averaging relied on our internal knowledge of which records belonged to the same individuals in each implicate. Since the public use implicates will not allow users to identify individuals across implicates, this is again an aggressive matching strategy. We found that match rates between the Gold Standard and the "average implicate" were higher than for the single implicate test, but still did not exceed 5%. This gives us a great deal of confidence that even if an intruder could match implicates perfectly, he or she still would not be highly successful in matching back to the SIPP public use files.

Table 1: Agreement Probabilities for Individuals with Spouses

| Field | Comparison Type | Pr(agree \| match): m | Pr(agree \| non-match): u | Agree weight: ln(m/u) | Disagree weight: ln(1-m)/(1-u) |
|---|---|---|---|---|---|
| Birthdate | D | 0.911727 | 0.000001 | 14.144163 | -2.427322 |
| Hispanic | c | 0.954479 | 0.835287 | 0.133390 | -1.286023 |
| Educ_5cat | c | 0.330004 | 0.241200 | 0.313478 | -0.124467 |
| Disab_in_scope | c | 0.949006 | 0.777256 | 0.199645 | -1.474307 |
| Disab | c | 0.843075 | 0.810676 | 0.039187 | -0.187691 |
| Disab_nowork | c | 0.637131 | 0.541970 | 0.161765 | -0.232893 |
| Totfam_kids_wave2 | c | 0.469601 | 0.329187 | 0.355257 | -0.234861 |
| Ind_4cat | c | 0.361122 | 0.309276 | 0.154980 | -0.078026 |
| Foreign_born | c | 0.844434 | 0.788724 | 0.068250 | -0.306097 |
| Time_arrive_usa | c | 0.236797 | 0.162303 | 0.377738 | -0.093133 |
| Ind_exist | c | 0.762450 | 0.568762 | 0.293074 | -0.596280 |
| Occ_exist | c | 0.775007 | 0.572171 | 0.303434 | -0.642654 |
| Occ_4cat | c | 0.446905 | 0.343057 | 0.264449 | -0.172067 |
| Mh_category | c | 0.591162 | 0.574111 | 0.029268 | -0.040861 |
| Flag_mar4t | c | 0.987294 | 0.987260 | 0.000035 | -0.002695 |
| Own_home | c | 0.719070 | 0.668007 | 0.073660 | -0.167008 |
| Pension_in_scope_age | c | 0.976252 | 0.949419 | 0.027870 | -0.756061 |
| Pension_in_scope_empl | c | 0.702327 | 0.557740 | 0.230506 | -0.395902 |

Table 2: Agreement Probabilities for Single Individuals

| Field | Comparison Type | Pr(agree \| match): m | Pr(agree \| non-match): u | Agree weight: ln(m/u) | Disagree weight: ln(1-m)/(1-u) |
|---|---|---|---|---|---|
| Birthdate | d | 0.872982 | 0.000001 | 13.664781 | -2.063423 |
| Hispanic | c | 0.888222 | 0.817697 | 0.082729 | -0.489153 |
| Educ_5cat | c | 0.360123 | 0.252198 | 0.356231 | -0.155862 |
| Disab_in_scope | c | 0.923310 | 0.744927 | 0.214679 | -1.201784 |
| Disab | c | 0.824805 | 0.113998 | 1.978968 | -1.620817 |
| Disab_nowork | c | 0.679595 | 0.222995 | 1.114350 | -0.885862 |
| Totfam_kids_wave2 | c | 0.568113 | 0.130233 | 1.472992 | -0.700061 |
| Ind_4cat | c | 0.356281 | 0.305685 | 0.153165 | -0.075664 |
| Foreign_born | c | 0.852712 | 0.094033 | 2.204775 | -1.816610 |
| Time_arrive_usa | c | 0.289757 | 0.091983 | 1.147440 | -0.245656 |
| Ind_exist | c | 0.784428 | 0.603121 | 0.262838 | -0.610339 |
| Occ_exist | c | 0.784490 | 0.602726 | 0.263572 | -0.611621 |
| Occ_4cat | c | 0.465897 | 0.388607 | 0.181394 | -0.135150 |
| Mh_category | c | 0.763459 | 0.067933 | 2.419334 | -1.371281 |
| Flag_mar4t | c | 0.990087 | 0.004855 | 5.317686 | -4.609064 |
| Own_home | c | 0.547307 | 0.242271 | 0.814954 | -0.515111 |
| Pension_in_scope_age | c | 0.887510 | 0.585350 | 0.416210 | -1.304568 |
| Pension_in_scope_empl | c | 0.693329 | 0.211577 | 1.186915 | -0.944258 |

Table 3: Match rates for Married and Single Individuals using Probablistic Record Linking

Married Individuals

| Type of Match | Total True | Total Records | Match Rate | Ratio of 2 to 1 | Ratio of 3,2 to 1 |
|---|---|---|---|---|---|
| Highest scoring | 4418 | 134662 | 0.0328 | 0.4033 | 0.7558 |
| Second highest scoring | 1782 | 134662 | 0.0132 | | |
| Third highest scoring | 1557 | 134662 | 0.0116 | | |

Single Individuals

| Type of Match | Total True | Total Records | Match Rate | Ratio of 2 to 1 | Ratio of 3,2 to 1 |
|---|---|---|---|---|---|
| Highest scoring | 4147 | 129132 | 0.0321 | 0.3053 | 0.5889 |
| Second highest scoring | 1266 | 129132 | 0.0098 | | |
| Third highest scoring | 1176 | 129132 | 0.0091 | | |

Table 4: Mahalanobis Distance Matching Results

| Male | Marital Status | N Synth | N N GS | Match Rate 1 Maha1 | Match Rate 2 Maha1 | Ratio 2 to 1 | Match Rate 3 Maha1 | Ratio 3 to 2 | Ratio 3, 2 to 1 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 70,814 | 70,814 | 1.11 | 0.50 | 0.45 | 0.44 | 0.88 | 0.84 |
| 0 | 1 | 70,478 | 70,478 | 1.03 | 0.55 | 0.53 | 0.44 | 0.81 | 0.96 |
| 1 | 4 | 39,434 | 39,434 | 0.97 | 0.52 | 0.54 | 0.39 | 0.74 | 0.93 |
| 0 | 4 | 34,481 | 34,481 | 1.18 | 0.73 | 0.62 | 0.55 | 0.74 | 1.09 |
| 0 | 3 | 18,733 | 18,733 | 1.05 | 0.54 | 0.51 | 0.33 | 0.61 | 0.83 |
| 0 | 2 | 14,668 | 14,668 | 1.04 | 0.67 | 0.64 | 0.50 | 0.74 | 1.12 |
| 1 | 3 | 12,370 | 12,370 | 1.04 | 0.46 | 0.44 | 0.38 | 0.82 | 0.81 |
| 1 | 2 | 2,815 | 2,815 | 2.91 | 1.53 | 0.52 | 0.78 | 0.51 | 0.79 |
| **Totals** | | 263,793 | 263,793 | 1.09 | 0.57 | 0.52 | 0.44 | 0.79 | 0.93 |

| Male | Marital Status | N Synth | N N GS | Match Rate 1 Maha2 | Match Rate 2 Maha2 | Ratio 2 to 1 | Match Rate 3 Maha2 | Ratio 3 to 2 | Ratio 3, 2 to 1 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 70,814 | 70,814 | 0.80 | 0.39 | 0.48 | 0.31 | 0.81 | 0.87 |
| 0 | 1 | 70,478 | 70,478 | 0.67 | 0.38 | 0.57 | 0.32 | 0.83 | 1.05 |
| 1 | 4 | 39,434 | 39,434 | 0.68 | 0.39 | 0.58 | 0.28 | 0.71 | 0.99 |
| 0 | 4 | 34,481 | 34,481 | 0.80 | 0.50 | 0.63 | 0.42 | 0.84 | 1.15 |
| 0 | 3 | 18,733 | 18,733 | 0.64 | 0.40 | 0.62 | 0.34 | 0.85 | 1.15 |
| 0 | 2 | 14,668 | 14,668 | 0.78 | 0.41 | 0.53 | 0.38 | 0.93 | 1.02 |
| 1 | 3 | 12,370 | 12,370 | 0.74 | 0.30 | 0.41 | 0.35 | 1.16 | 0.88 |
| 1 | 2 | 2,815 | 2,815 | 2.20 | 0.99 | 0.45 | 0.75 | 0.75 | 0.79 |
| **Totals** | | 263,793 | 263,793 | 0.75 | 0.41 | 0.55 | 0.34 | 0.83 | 1.00 |

Table 5: Euclidean Distance Matching Results

| Male | Marital Status | N Synth | N N GS | Match Rate 1 EUCL1 | Match Rate 2 EUCL1 | Ratio 2 to 1 | Match Rate 3 EUCL1 | Ratio 3 to 2 | Ratio 3, 2 to 1 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 70,814 | 70,814 | 0.60 | 0.40 | 0.66 | 0.31 | 0.77 | 1.17 |
| 0 | 1 | 70,478 | 70,478 | 0.58 | 0.39 | 0.67 | 0.27 | 0.71 | 1.15 |
| 1 | 4 | 39,434 | 39,434 | 0.49 | 0.28 | 0.58 | 0.21 | 0.75 | 1.01 |
| 0 | 4 | 34,481 | 34,481 | 0.53 | 0.32 | 0.61 | 0.30 | 0.93 | 1.18 |
| 0 | 3 | 18,733 | 18,733 | 0.90 | 0.57 | 0.63 | 0.36 | 0.63 | 1.03 |
| 0 | 2 | 14,668 | 14,668 | 0.47 | 0.42 | 0.90 | 0.22 | 0.53 | 1.38 |
| 1 | 3 | 12,370 | 12,370 | 0.74 | 0.45 | 0.61 | 0.40 | 0.88 | 1.14 |
| 1 | 2 | 2,815 | 2,815 | 0.82 | 0.50 | 0.61 | 0.36 | 0.71 | 1.04 |
| **Totals** | | 263,793 | 263,793 | 0.59 | 0.38 | 0.65 | 0.29 | 0.75 | 1.14 |

| Male | Marital Status | N Synth | N N GS | Match Rate 1 EUCL2 | Match Rate 2 EUCL2 | Ratio 2 to 1 | Match Rate 3 EUCL2 | Ratio 3 to 2 | Ratio 3, 2 to 1 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 70,814 | 70,814 | 1.26 | 0.74 | 0.58 | 0.55 | 0.75 | 1.02 |
| 0 | 1 | 70,478 | 70,478 | 1.43 | 0.81 | 0.57 | 0.66 | 0.81 | 1.03 |
| 1 | 4 | 39,434 | 39,434 | 0.94 | 0.59 | 0.62 | 0.51 | 0.87 | 1.16 |
| 0 | 4 | 34,481 | 34,481 | 1.16 | 0.67 | 0.58 | 0.51 | 0.76 | 1.02 |
| 0 | 3 | 18,733 | 18,733 | 0.91 | 0.56 | 0.61 | 0.42 | 0.76 | 1.07 |
| 0 | 2 | 14,668 | 14,668 | 1.03 | 0.53 | 0.52 | 0.52 | 0.99 | 1.03 |
| 1 | 3 | 12,370 | 12,370 | 0.91 | 0.53 | 0.58 | 0.44 | 0.85 | 1.06 |
| 1 | 2 | 2,815 | 2,815 | 2.31 | 1.17 | 0.51 | 1.03 | 0.88 | 0.95 |
| **Totals** | | 263,793 | 263,793 | 1.20 | 0.70 | 0.58 | 0.56 | 0.81 | 1.05 |