Matching Survey of Occupational Injuries and Illnesses Frame Data to Occupational Safety and Health Administration ITA Data

Task Order: 1625DC-17-A-0002

Final Report

JUNE 28, 2019

PRESENTED BY:

Stephen Cohen, Senior Fellow

Dean Resnick, Senior Data Scientist

Evan Herring-Nathan, Statistician

NORC at the University of Chicago

Contents

Linking Methodology

The approach used

As proposed in the methodology report, we used a modification of the methods proposed by Fellegi and Sunter (Fellegi & Sunter, 1969; Resnick, 2017) to link the 2016 OSHA ITA records to the records on the 2016 SOII frame. The records were scored and evaluated according to the level of agreement of the shared identifiers through the estimation of match probabilities and agreement (and non-agreement) weights. Fellegi and Sunter demonstrated that linkage accuracy (minimizing Type I error for given level of Type II error) can be optimized by setting these weights according to the ratio of agreement probabilities (and their complements) among matched pairs (i.e., those representing the same entity, in the context of this linkage, the same establishment (called M-probabilities) and unmatched pairs (called U-probabilities). We found this to be true and that the method was viable.

The machine-learning algorithm that we used estimates M- and U- probabilities in the absence of training data (data in which true match status is known for a sufficiently large number of pairs such that reliable matching parameters can be estimated). While the level of fit convergence depends on the properties of the data we developed appropriate settings that produced good results. We utilized 10 non-continuous identifier agreement variables to realize a model with good-fit to the data.

The essence of this fitting approach is as follows. The candidate pairs to be evaluated are categorized and tabulated by their agreement pattern. The agreement pattern is a vector of the pairwise agreement status of each of the comparison variables, as firm name, ZIP, census block, NAICS code, etc. Each of the comparison variables are evaluated to either agree (represented by the value 1), (disagree represented by the value 0), or be indeterminate if values are not available to compare (as would by the case for firm name if it was not reported). Pairs are either matches (i.e., comprised records represent the same firm) or are non-matches, and we assume that for each of them for each comparison variable there are uniform probabilities of agreement. For example, for matches, the probability that NAICS, first two-digits agree is 95%.

Thus it is possible to estimate both the probability that a matched record and the probability of an unmatched records has a certain agreement pattern, and if the number of matches and non-matches is known we can estimate the number of pairs we expect to have each possible agreement pattern. The comparison variable agreement rates (for each one, one for matches and one for non-matches) and the total number of matches are parameters then which when adjusted yield various estimates of the number of pairs for each agreement pattern. By comparing these expected counts to the actual counts we can compute a goodness-of-fit statistic. The fitting algorithm then seeks to minimize this statistic. The minimum is found by an Newton-Raphson search methodology starting from initial guesses—each iteration of the search reduces (improves) the goodness-of-fit statistic until no more improvements can be found.

The parameter estimates found using this machine learning algorithm are generally consistent with methods proposed by Fellegi and Sunter. However, we make appropriate modifications for handling identifier-agreement dependence. While, the traditional Fellegi-Sunter pair scoring assumes that identifier agreements are conditionally independent—that is to say that agreement of one identifier has no bearing on the agreement status of another—we expect that for some identifiers, the assumption of conditional independence is unreasonable.

To account for this dependence, we use odds-adjusters to modify the unconditioned odds of agreement such that the conditioned odds is equal to the product of the odds adjuster and the unconditional odds. The odds adjusters themselves are treated as parameters in the fitting process and the optimized values are realized through iteration and convergence. For more details see Resnick (2017).

Also, the use of nested agreement identifiers (e.g., we can only have agreement on all six digits of NAICS if we have agreement on the first two digits of NAICS) necessitate a revamping of the probabilistic assumptions holding in the Fellegi-Sunter paradigm. We have applied these revamped assumptions as an experimental feature of our machine learning algorithm with successful results. They require that the nesting relations be explicitly coded within the parameter settings.

Summary of the steps involved in the algorithmic approach

In the process of linking the ITA and SOII frame data many decisions were made to optimize the approach. Initially, we developed a solid understanding of the input data, including how the data were collected and what the fields represent, was developed. In consultation with BLS staff, NORC was able to understand these data and implement the algorithm such that data were correctly utilized.

Decision points and BLS input

One specific area where initially there was a lack of clarity was in the concept of addresses and geographical considerations. First, consensus was reached on what an establishment is and following from that, what types of pairs can be considered to be true establishment matches. An establishment was defined as a specific physical location (e.g. store front, floor of an office building, warehouse) as opposed to a business or division of a business which includes establishments at multiple locations. Only pairs that are believed to reflect the same physical location (i.e., represent the same establishment) would be considered true matches. While the ITA data only had one address to use, the SOII frame had three: Physical, Unemployment Insurance, and Other. In consultation with BLS, it was determined that that in general the best representation of the location of an establishment is the physical address; however, in cases where the physical address (“PH” address fields) was non-geocodable, we would seek to replace it with the Other address (“MO” address fields). However if both the physical and other address were geocodable, we would use the MO address only in an ancillary linkage run whose predicted match probabilities would be understood to be contingent on the MO address actually representing the physical location. For more details on the treatment of the address fields refer to the User’s Manual.

The ITA establishments are a subset of all establishments in the SOII frame. As a result, there are two ways in which to view the links; from an OSHA-centric or a SOII-centric perspective. With guidance from BLS, an OSHA-centric perspective was taken wherein for each ITA record only the highest scoring pair was retained in the final link file. This approach keeps file size manageable as the ITA file consists of ~217,000 records as opposed to SOII frame’s ~8 million. It also is appropriate since only a subset of the SOII frame is required to report to OSHA and there is both over- and under-compliance. Because all records are retained, and no specific match probabilities thresholds are used, BLS has maximum flexibility in determining which pairs are true matches. They are also able to analyze other characteristics of the linkage.

To quantify the level of agreement between records' business names several considerations were taken. First, both fields in the SOII frame related to the business' name (company name and establishment name) were used. Often the two name fields are identical or the establishment name is not meaningful. But there are situations where a blank company name was imputed with its establishment name field to improve match results. Sufficient standardization on the name fields, as well as other fields were performed to deal with data quality issues.

Contact information, email addresses and telephone numbers of the business, was also used in the linkage and merged on to the ITA and SOII frame files. The contact information was not unique to the record level and it created records with more than one email address and/or telephone number. Once the file was scored, for each ITA-SOII frame pair combination, only the highest scoring pair was retained. This step ensured that all available contact information was used effectively.

Problems encountered

Computational requirements

The computational requirements of this record linkage are a critical consideration. The amount of data that must be processed is significant and the computations are intensive. Even with the use of zip code blocking there were approximately 170 million pairs to score for the PH address linkage. These characteristics make performing the linkage on ordinary computers impossible. We used a Linux server with 2TB of storage and high computing power to run our SAS programs. We found that this server proved sufficient for our needs and we assume that future iterations of the linkage would be runnable with a similar setup. While the creation and scoring of the pairs files was performed on the Linux server, the geocoding process, where addresses are standardized and matched to the Census TIGER files, was performed on the user's local machine because speed was considerably faster than on the Linux server. For more details on computational requirements and settings refer to the User's Manual.

Convergence

The largest difficulty with using our linkage algorithm is the possibility of non-convergence. For this analysis, we found that we got substantially better linkage results by estimating parameters with a random subset of pairs (i.e., we used 1%) rather than with the entire set, although it is not clear to us why this is the case. One issue with running the fitting algorithm on a subsample is that certain levels of the comparison vector which exist among the full set of pairs are not included in the sample and so do not generate a corresponding probability of selection. We resolved this issue by developing code which uses the estimated parameters to predict match likelihood for the left out agreement patterns in the comparison vector.

More generally, we experimented with different settings for the algorithm parameters to find the one which resulted in the best results. The specific comparison variables used, inclusion of interaction terms, the starting value for the estimated number of matches, and the number of iterations to perform all had an impact in the quality of the algorithm's convergence and linkage results. This type of experimentation is considered to be typical in the application of the algorithm and does not reflect a problem in the technical sense but it does warrant attention from users of the algorithm.

As an alternative to the fitting routine which we have used for the application of our algorithm, we are also providing a SAS code module that allows the M-, and U- parameters to be estimated using the more standard expectation-maximization algorithm. If comparison variable independence is assumed, the E-M algorithm parameter estimates can be used to estimate match likelihood for all levels of the comparison vector. Alternatively, rather than using an estimate of match likelihood the actual pair comparison score can be used to rank match likelihood. The comparison score is found by summing over all comparisons either the M weight (when there is agreement) or the U weight (when there is non-agreement).

Agreement vector and comparison variables

To maximize the usefulness of the information contained in the data, often one variable was used to create multiple agreement indicators. For example, after investigating the data we found that the NAICS code could most effectively be used by considering both the 2-digit NAICS and the 6-digit NAICS. The 4-digit NAICS was not especially useful because the agreement status that resulted was not significantly different from that of the 6-digit version (i.e. when there was 4-digit agreement there was often 6-digit agreement). And so, to reduce complexity of the model, and improve fit characteristics, the simpler and more efficient configuration was used.

The business name fields were also used in multiple ways to be able to differentiate between pairs with varying levels of name agreement. After experimentation with various configurations we found that a 3-tiered approach was effective. First the pairs' business names were tokenized. The process of tokenization involves parsing the business name field into its component words. Each non-trivial token (i.e. word that is longer than 3 characters) of the business name in ITA was compared to each non-trivial token in the SOII frame. The highest-level of agreement (most basic agreement) required that there be at least one non-trivial token in agreement. The next level, nested in the highest level, requires that the pair has an organization name agreement score of at least 10,000 according to the business name matching algorithm. As described in the methodology report, the token comparison score is incremented proportional to the rarity of the token in agreement. The most discerning agreement indicator required very high agreement characteristics. The pair had to satisfy one of the following:

- Agreement on at least 50% of the tokens (relative to the record with the shorter name field) and has a score of at least 30,000
- Has a score of 20,000 with token count agreement greater than 1

A careful review of all of the available data elements along with input from BLS data user's with in-depth knowledge of the SOII frame data and BLS' research of businesses in general was important in informing the linkage. Additionally, background information on the ITA data including a data dictionary were crucial in determining how to quantify agreement and match probability. All variables that were available and appropriate for use were considered for the algorithm. Experimentation was required to find the right mix of agreement indicators; not only which variables were used but how they were used (e.g. levels of geographical proximity and business name agreement). For details on the agreement indicators used refer to the User's Manual.

In addition to experimentation with the data inputs to the algorithm, consideration was also given to the model parameter settings. Multiple attempts were performed and results were analyzed to improve and determine the optimal settings. The process is somewhat slow in the sense that the algorithm takes some time to run and there must be some amount of manual review to determine fit. However, the most time-intensive part of the linkage process is in the geocoding of the addresses and the creation of the pairs. These processes are essentially automated provided that the user has supplied the right values for the macro variables. They only require an appropriate amount of review to determine that the results are as expected. From there the data can be scored and a sample of records can be taken from various match probability ranges to determine if model match probability matches user expectations and estimates. Details on this process are found in the User's Manual.

Establishments that have a duplicative address but are distinct establishments bring unique challenges to an accurate linkage. Examples of this are found at airports where multiple airlines have the same address. This also occurs at strip malls and office buildings with multiple suites. A multi-unit indicator was created using the SOII frame where addresses that appeared twice but did not seem to be representing the same business (e.g. a duplicate record) were flagged. Ultimately this indicator was not used directly in the linkage inputs but can be used after the linkage is preformed to inform the manual determination of match status. The reason that this indicator was not used in the process is due to the fact that it didn't improve the fit characteristics and we were not able to determine how to otherwise appropriately adjust for this variable in the computation of match probability. More experimentation with this concept could be justified for future iterations.

Blocking

Consideration to the geographic location of the establishments was instrumental in setting the blocking scheme and in making the file size manageable. Even with the use of blocking, considerable computational resources were required to process the input data and create the links. BLS and NORC agreed to use zip code as the blocking factor. This decision was motivated by a desire to capture as many true matches as possible while creating pair files that would be tractable with available resources. A manual review of a sample of links found that many matches would not have been found when blocking by Census block and so using zip code was preferable. Zip code is also a good choice since establishments that do not agree on zip code can't represent the same physical establishment, except in cases of data quality issues. However, addresses from different blocks could represent the same establishment (e.g. receiving warehouse adjacent to store front). Since addresses were standardized and data quality is believed to be high, our approach is sensible.

Description of datasets

There are four input datasets used in the process. The OSHA ITA dataset from 2016 contained ~217,000 records and includes information on the businesses who reported as well as the collected data related to injuries that BLS seeks to add to the existing SOII frame. The SOII frame consists of over 8 million businesses from 2016 and contains information about the businesses (name, location, business type, industry, etc.)

Some contact information was provided on separate supplemental files, one for the SOII frame and one for ITA. They include, when available, an email address and telephone number for the user associated with the record. The percentage of records on these supplemental files that have either an email address or telephone number is small and the percentage that have both is even smaller. When combined with the telephone number available on the SOII frame, these variables did provide some benefit to the algorithm and were used.

The Census TIGER files are also a required input to the process. These files contain the street lookup data from which geographic location of the businesses can be determined. They help to standardize the street address to enable a more accurate match and also help establish agreement at the block, tract, zip code, and county levels.

Description of the linking software

How to use the software

To run the linkage programs the user must first read through the instructions in the User's Manual. The first of these steps is to source the input data (SOII frame, ITA, contact information for both, and the SAS TIGER files for geocoding). The data must then be processed to perform standardization and geocoding in preparation for the linkage. Then pairs are formed by blocking on zip code resulting in a very large number of pairs.

Agreement indicators are set according to the agreement status of the selected variables. Then the model is fit and results are analyzed by the user. If fit is determined to be unacceptable the user makes adjustments to the selected variables and iterates through the process until a satisfactory model is realized. If a model can't be realized the user may use alternative methods mentioned in the methodology report. Once a method is chosen, the data are scored and records are classified as being either near certain matches, near certain non-matches, or pairs requiring further review.

The process has been automated to the extent possible and macro programs are utilized throughout. By supplying compatible data, with similar structure and characteristics as the 2016 data that was used, and then updating macro parameters appropriately, the user can create links for new data. The process is complex and requires many steps to be taken in the right order to successfully complete. The process is broken up into several SAS programs and separate macros to improve readability of the code.

Many of the steps also require significant computational time and resources and interaction on the part of the user even with an adequate understanding of the process. These steps and considerations are outlined in the User's Manual which will guide the user in the successful completion of the linkage. The manual also provides general information on how to use the programs that will be critical in user understanding. The attached SAS code is also commented to guide the user in the correct implementation.

Description of the linked file on the BLS server

The linked files

The linkage code produces two linkage files. The main linkage file is based primarily on the SOII frame physical (PH) address as described above and in the User's Manual. The second linkage file is the alternative version that uses the SOII frame other (MO) address instead of the PH address. In both files the ITA address used is the only one available in the ITA file.

Both of the files contain one and only one pair for each ITA record on file. Records on ITA that could not be paired to any SOII frame records in the blocking process are not included (~200). Since ITA records can be paired with multiple SOII frame records due to the blocking scheme, a duplication step was performed to include only the highest scoring pair for each ITA record. The match probability for the pair is included as a variable in the output file and it can be used to determine true match status and for other analyses. Other variables included are the various ID fields from the data to be able to match businesses back to their source data and to other data as well. The variables that are used to create the agreement indicators, as well as the agreement indicators themselves, are included in the output files. Collectively, the provided data elements allow the user to assess model fit, to perform subsequent analysis, and to estimate information collected on ITA for establishments on the SOII frame.

Accuracy measures

We have evaluated linkage accuracy in several ways. We ensured the viability of the model by considering the model fit characteristics. Convergence in the algorithm was observed as the fit score ratio was at an acceptable level. Model parameters seemed to be in line with expectations given the quality of the data. An analysis of the agreement patterns and corresponding match probabilities suggests that results are plausible. For example, pairs that have agreement on all selected fields, while fewer in number have a virtually 100% probability of being a true match. Pairs with conflicting geographic similarity (different Census block or address but same zip code) and a similar business name are flagged as having a possible, but not an almost certain, probability of being a true match.

A random sample of pairs at various match probability levels suggested that model estimated probabilities were similar to user-estimated probabilities. Even a manual review cannot definitively classify all pairs as matches or non-matches due to the unique characteristics of these data and the challenges in linking business establishments. The probabilities are then useful to make effective judgments about match status and accuracy trade-offs. The setting of the probability thresholds is left to BLS so that customized approaches may be used.

While not currently available, the development of a truth deck could offer additional insight into the accuracy of the linkage. The truth deck would contain known match status for a certain number of pairs and would include the same data elements as the pairs used in developing the model parameters and pair match probabilities. Then the developed model could be applied to the truth deck and the true match status can be compared with the estimated match probability. If the group of pairs with estimated match probabilities above 95%, for example, have an actual true match rate of ~95% then the model fit would be very good. This approach would require considerable effort in the creation of the truth deck. A sufficient number of true matches for each agreement pattern would need to be accurately produced in order to apply this method.

Perhaps a more efficient way to estimate linkage accuracy is to perform the sampling method described above. A small random sample of matched pairs and unmatched pairs is taken being sure to include all levels of match probabilities. These pairs are then grouped into a certain number of classes according to their probabilities (perhaps 10 classes evenly spaced). Then intensive manual review can determine what percent of each set were classified correctly by the linkage. BLS is encouraged to perform a comprehensive review of this nature to the extent desired to allow for their independent evaluation of linkage accuracy and to promote their confidence in the methodology used. While NORC has performed this sampling evaluation to some degree, it is cost-prohibitive for NORC to engage in a comprehensive clerical-review and our proposal noted this limitation.

Assuming that model fit is good and that estimated match probabilities are accurate it is then possible to develop error rates. For example, for an agreement pattern with a match probability of x% that is above the threshold for being considered a true match (i.e., the linkage threshold) it follows that (1-x)% of those matches are false positives. The overall false positive rate can then be derived by considering the rate for each agreement pattern and the number of matches having that agreement pattern. Likewise, for agreement patterns which are below the linkage threshold (and so, not linked), we can estimate that x% of these are false negatives. Multiplying this percentage by the number of pairs with this pattern estimates the false negatives with this pattern, and summing over all of these patterns below the linkage threshold estimates the total number of false negatives: i.e., true matches that are not linked.

Both the evaluation methods performed above as well as the establishment of a truth deck should be completed by BLS in future iterations of the linkage to help determine linkage accuracy.

Future considerations on application of the software

Strengths and weaknesses of the algorithm

The approach that we have developed is effective in matching businesses and differentiating probable matches from probable non-matches. We believe it to be the best approach given the available data and resources and given the absence of a truth deck. The estimated match probabilities provide confidence in knowing that identified links are most likely true and they allow the user to set customized thresholds depending on the desired precision and specificity. The model specification is flexible and allows different parameter settings to customize the method for new data. There are several linkage accuracy evaluation methods available and results can easily be analyzed to determine if the estimated match probabilities are plausible.

True of any approach, our approach is only as strong as the available data, which is observed to have some degree of data quality issues. By improving input data quality accuracy will improve. Also, like other approaches the process cannot be fully automated. Manual review is required in the development of truth deck and in the evaluation of the links.

Due to the volume of data the process is computational expensive. Since powerful resources are available this does not represent a barrier to linking the data but it does mean that sufficient time and resources need to be given to the process. It's probable that our approach requires less manual effort than other approaches, and it certainly reduces effort by blocking on zip code. Further efficiencies are realized by considering only the ITA records' highest scoring pair.

Barring big changes to the nature of BLS's SOII frame data and OSHA's ITA data, how they are collected, and the available data elements, we expect that the linkage will work for future years' data without significant modification to the code. The use of appropriately organized macro programs along with detailed instructions make the process of modification much more intuitive.

While we have attempted to evaluate linkage accuracy to the best degree given available resources, we have only sampled a portion of the file. A more thorough review of all of the pairs can provide more insight and detailed understanding of the accuracy. As is the case with all linkage approaches, some matches will be missed while others will be incorrectly identified. We believe that our approach sufficiently balances these tradeoffs.

The process may be improved by incorporating experts in the SOII frame and ITA data. Including these people at certain stages of the process may inform the approaches taken and improve the match. Review by someone more familiar with business data may help improve the evaluation.

The algorithm does an adequate job of differentiating different business that have the same physical address. However, additional methods to ensure that the presence of these pairs do not compromise accuracy may be helpful.

What BLS should be considering when moving forward

As mentioned, BLS may be advised to develop a truth deck. With enough records that are known to be true matches the parameters could more accurately be estimated and it wouldn't be necessary to run the algorithm. Also, the truth deck could be used to evaluate linkage accuracy.

BLS can also explore to see if there is additional data that could be used to create even more agreement indicators. Given the availability of ample identifier variables, a hold-out variable could be used to determine the relative level of agreement on this variable for matches as compared to non-matches. If the matches display significantly higher agreement than the non-matches this will provide evidence that the match is accurate and will be useful in the quantification of the accuracy. If agreement for matches is similar to non-matches the evidence will be equivocal since it may not be possible to determine if the similarity is due to lack of utility in the hold-out variable or an actual limitation in linkage process. If agreement for matches is lower than non-matches, this will be clear indication that the linkage was not successful.

The exact process of the hold-out approach is described in more detail. A linkage is performed without the use of the hold-out variable. Then the agreement rate on the hold-out variable is used to estimate the proportion of links that are valid. For example, if we do linkage on name and address but exclude industry group, then we can use the rate of agreement on industry group to estimate the proportion of links that are true matches. Imagine that it was determined during the fitting process that the M-probability (agreement for matched pairs) for industry is .95 and its U-probability (agreement for unmatched pairs) is .1. Then the linked data reveals that when name and address agree industry agrees 80% of the time. With this information we can use algebra to estimate the ratio of matches to non-matches among pairs agreeing on name and address through the equivalence of 80% = (Matches x .95 + Non-Matches x .10) / (Total Pairs) subject to the constraint that the sum of the number of matches and non-matches equals the denominator (Total Pairs).

Conclusion

NORC has developed a complete process including a machine learning algorithm for linking the OSHA ITA data to the SOII frame. We have provided SAS code for performing all steps in the process and have also included a User's Manual to assist in the understanding and application of the code. We applied this process to the 2016 data sources and created two link files containing the pairs with the highest probability of being a match for each ITA record that was able to be blocked. The resources provided by NORC should be sufficient for BLS in future applications of the code and for making adjustments to handle new characteristics of future years' data.

Appendices

References

Fellegi, Ivan P., and Alan B. Sunter. "A theory for record linkage." *Journal of the American Statistical Association,* 64.328 (1969): 1183-1210.

Resnick, Dean M. "The Estimation of Match Validity under the Fellegi-Sunter Paradigm without Assuming Identifier-Agreement Independence," *Proceeding of the Joint Statistical Meetings*, 2017.